DOCUMENT RESUME

ED 148 890 TM 006 827

AUTHOR Reckase, Mark D.
TITLE Ability Estimation and Item Calibration Using the One and Three Parameter Logistic Models: A Comparative Study. Research Report 77-1.
INSTITUTION Missouri Univ., Columbia. Tailored Testing Research Lab.
SPONS AGENCY Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office.
PUB DATE Nov 77
CONTRACT N00014-77-C-0097
NOTE 80p.

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS *Achievement Tests; Aptitude Tests; *Comparative Statistics; *Cost Effectiveness; Data Analysis; Goodness of Fit; Group Tests; *Item Analysis; Item Banks; *Mathematical Models; Sampling; Simulation; Statistical Analysis; Testing Programs; Test Items
IDENTIFIERS *Latent Trait Models; *Tailored Testing

ABSTRACT

Latent trait model calibration procedures were used on data obtained from a group testing program. The one-parameter model of Wright and Panchapakesan and the three-parameter logistic model of Wingersky, Wood, and Lord were selected for comparison. These models and their corresponding estimation procedures were compared, using actual and simulated test results, with respect to: (1) the ability to calibrate multivariate data, (2) the sample size needed for calibration, (3) the effects of item quality on the calibration, and (4) the operational costs. The one-parameter procedure provided equivalent ability estimates at a lower cost than the three-parameter procedure when basically unifactor group tests were used. This method was therefore recommended for item calibration of group administered, multiple choice tests. However, the goodness of fit of the models to the data definitely showed the three-parameter model to be superior. Whether the better fit will outweigh the higher costs can only be determined by further comparison of the usefulness of the ability estimates, in actual tailored testing. (Author/MV)

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Research Report 77-1	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Ability Estimation and Item Calibration Using The One and Three Parameter Logistic Models: A Comparative Study		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Mark D. Reckase		8. CONTRACT OR GRANT NUMBER(s) N00014-77-C-0097
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Educational Psychology University of Missouri Columbia, Missouri 65201		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P.E.: 61153N Proj.: RR042-04 T.A.: 042-04-01 W.V.: NR150-395
11. CONTROLLING OFFICE NAME AND ADDRESS Personnel and Training Research Programs Office of Naval Research Arlington, Virginia 22217		12. REPORT DATE October, 1977 13. NUMBER OF PAGES 66
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approval for public release; distribution unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Testing	Rasch Model
Ability Testing	Multi-factor Tests
Latent Trait Models	Tailored Testing
	Achievement Testing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The literature on latent trait calibration procedures was reviewed to determine the methods available to calibrate dichotomous items for tailored testing applications. From the procedures, the most promising techniques for the calibration of items using the one- and three-parameter logistic models were selected for comparison. The maximum likelihood procedure developed by Wright & Panchapakesan was selected for the one-parameter model; and the estimation procedure for use with omitted responses,

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

#20 (cont)

developed by Wood, Wingersky & Lord, was selected for the three-parameter model. The two procedures were then compared on their suitability for use with multivariate item pools, the sample size required for calibration, the effects of item quality, and the cost of calibration. Sixteen data-sets were used for these evaluations; eight live testing data-sets, and eight simulation data-sets generated to match specified factor structures. The one- and three-parameter models were found to estimate different components when the tests measured several independent factors. The three-parameter model estimated parameters from one of the factors, ignoring the others, while the one-parameter model estimated the sum of the factors. When a dominant first factor was present in the test, the two calibration procedures calibrated the items and estimated ability on the first factor. Although the three-parameter model fit the test data significantly better than the one-parameter model, there was no difference for the two procedures in predicting outside criterion measures. The sample size analysis showed that the one-parameter model required substantially fewer cases for item calibration than the three-parameter model. Some general sample size recommendations were made. Item quality, as determined by guessing and discrimination parameter estimates, was found to affect the fit of the two models to the data. However, the probability of fit statistic given by the one-parameter logistic program was affected only by guessing. In terms of cost, the three-parameter procedure was found to be substantially more expensive than the one parameter procedure. Although the three-parameter model was found to fit the data better than the one-parameter model, the ability estimates from the two procedures were highly correlated when a dominant first factor was present; and the correlations with outside criterion measures were not significantly different. Since the one-parameter model costs less to use, it is the recommended procedure for calibration of 50 item, group administered, multiple-choice exams. This recommendation does not generalize to tailored testing administration, but only to item calibration for group tests.

Ability Estimation and Item Calibration Using the One and Three Parameter Logistic Models: A Comparative Study

Latent trait measurement models have slowly made inroads into the applied areas of testing. Information functions based on latent trait theory have been used to construct aptitude tests (Marco, 1976) and the simple logistic model has been used to scale an achievement test (Woodcock, 1973), while major use of the models has been made in the area of tailored testing (Jensema, 1974; Lord, 1970; Reckase, 1974; Samejima, 1975). Despite the acceptance of the models, debate still exists concerning the relative value of the various types of models being used. The major facet of the debate is the number of parameters required to adequately describe empirical item characteristic curves. One point of view specifies that three parameters are required to describe the interaction of a person and an item: difficulty, discrimination, and guessing. The opposite position is that only one parameter, difficulty, is required.

Until a point was reached that the latent trait models were regularly applied to live testing situations, it was sufficient to let the debate continue on theoretical grounds with the clear edge to three parameter models when multiple choice items were being considered. However, with the increasing use of these models in applied settings, and with the lack of comparative studies, the need for direct empirical comparisons is clearly indicated. An evaluation is of special importance considering the needs of tailored testing, where speed of convergence to an ability estimate and computational efficiency are of great importance. Because of these added constraints, the simplicity of the one parameter latent trait models tends to balance the theoretical completeness of the three parameter models.

The general orientation of the research program guiding this study is to apply tailored testing to achievement measurement. This fact places a number of other constraints on the type of comparison required among the latent trait test models. Achievement tests typically are constructed based on different criteria than aptitude tests. Content validity is the desired goal, rather than the measurement of some set of unidimensional traits. Construction methods fit items to a table of specifications yielding tests that may be of substantial factorial complexity. Thus, sensitivity of the models to multifactor data is a major consideration.

A second factor that will influence the use of latent trait models for achievement testing is the size of sample required for item

calibration. If tailored tests are to be used in educational programs that have some modicum of flexibility, large samples of students responding to the items may not be attainable before modifications in the instruction make the test obsolete. Thus procedures that yield stable calibration results with relatively small samples will have an edge in terms of applicability. The fluidity of educational programs has a further effect on the qualities desired in the latent trait models. Often, because of the short time available for the construction of tests, the items in a test may not be of highest quality. Therefore, a good model for achievement testing should be able to function using mediocre items.

The purpose of the research reported here is to evaluate the one and three parameter logistic models for use in calibrating achievement tests for use in tailored testing. Toward that end the factors mentioned above (computational efficiency, robustness to multidimensionality, effects of sample size, effects of item quality) will be manipulated in comparing the models. However, the relevant literature will be reviewed before describing the research design in detail.

Review of the Literature

The literature on latent trait theory has mushroomed over the last several years. A count of references since 1974 has yielded well over one hundred entries. Since other good reviews of the general area are already available (i.e. Hambleton, Swaminathan, Cook, Eignor, and Gifford, 1977) this paper will not attempt to summarize the total research effort, but will be limited to the areas directly related to applying latent trait models to achievement tests. More specifically, the review will concentrate on the available item calibration procedures, the effects of violating the assumptions of the models, and the types of tests and sample sizes appropriate for analysis.

Item Calibration Procedures

Numerous methods have been developed to estimate the item and ability parameters of the latent trait models. These vary in sophistication and computational complexity from the early graphic methods used by Rasch (1960) to the conditional (Andersen, 1973) and unconditional (Wright & Panchapakesan, 1969) maximum likelihood, least squares (Brooks, 1964), and empirical Bayes point estimates (Meredith & Kearns, 1973) currently being used. Many approximation techniques have also been developed to reduce the complexity of computation and computer time required.

Of the many methods available, only those appropriate to the simple logistic and three parameter logistic models applied to dichotomously scored items will be presented here. Multivariate models, and those appropriate for nominal, graded, and continuous response data, will be included only when they apply to the specific models of interest.

Simple Logistic Procedures

A procedure for estimating the parameters of the simple logistic model was first presented in Georg Rasch's original exposition of the model (Rasch, 1960). His procedure takes advantage of the local independence and sufficient statistic properties of the model to independently estimate the ability and easiness parameters. The basic procedure follows.

The simple logistic model as presented by Rasch (1960) is given by

$$P\{x_{ij} = 1\} = \frac{A_i E_j}{1 + A_i E_j} \tag{1}$$

and

$$P\{x_{ij} = 0\} = \frac{1}{1 + A_i E_j}$$

where x_{ij} is the score on the item, A_i is the ability of Person i and E_j is the easiness of Item j. The logarithm of the ratio of the probability of a correct response to the probability of an incorrect response is called the logit and is given by

$$l_{ij} = \ln\frac{P\{x_{ij} = 1\}}{P\{x_{ij} = 0\}} = \ln A_i + \ln E_j. \tag{2}$$

If a second item, k, is given to Person i, the logit is

$$l_{ik} = \ln A_i + \ln E_k. \tag{3}$$

The difference between Equations 2 and 3 gives

$$l_{ij} - l_{ik} = \ln A_i + \ln E_j - \ln A_i - \ln E_k = \ln E_j - \ln E_k$$

which does not contain the ability parameter. The average logit for Person i over all of the items on the test is given by

$$\frac{\sum_{j=1}^{n} l_{ij}}{N} = l_{i\cdot} = \frac{\sum_j (\ln A_i + \ln E_j)}{N} = \ln A_i + \frac{\sum_j \ln E_j}{N} = \ln A_j + \ln E_\cdot. \tag{4}$$

Subtracting Equation 4 from Equation 2 gives the basic estimation equation for the procedure

$$l_{ij} - l_{i\cdot} = \ln E_j - \ln E_\cdot . \quad (5)$$

If the average log easiness is set equal to zero, this equation simplifies to

$$l_{ij} - l_{i\cdot} = \ln E_j . \quad (6)$$

Thus, the easiness parameters can be estimated from the difference between the logit for an ability level and an item, and the average logit over items. The estimated easiness parameter should be the same regardless of the ability level used for the procedure.

In order to improve the estimates obtained, Rasch takes advantage of the fact that Equation 6 is a linear equation of slope 1.0 between the two logit variables. Therefore, he plots the l_{ij} value against $l_{i\cdot}$ across ability levels and then fits a slope 1.0 line to the resulting scatter plot using an "eyeball" technique. The intercept of the plotted line is used as the easiness estimate for the item. A similar procedure is used for the ability parameter, except that the average logit over ability levels is used.

This procedure obviously yields only rough approximations to the true parameter values and, since much of the results are based on a subjective fit to a scatter plot, a fully computerized procedure is not possible. For these reasons, Rasch's procedure was used only in early exploratory studies of the model.

In 1964, Brooks modified Rasch's procedure to increase its objectivity. Instead of visually fitting a slope 1.0 line to the plot of specific ability group logits against average logits, he used linear regression procedures to fit a line using least squares methodology. This allowed a quasi-statistical test for goodness of fit of the model to the test data by testing the empirically obtained slope against the theoretical value of 1.0. Brooks admitted that this significance test was not precise because the sampling distribution of the slope obtained under these circumstances was unknown, but he felt it was better than the visual check of the slope that Rasch used. As in the previous technique, the intercept of the fitted line was used to obtain the parameter estimate.

Although Brooks' procedure was an improvement over Rasch's original demonstration technique, neither it, nor the simple logistic model itself gained much prominence until later in the decade. By that time however, a more sophisticated maximum likelihood procedure had been derived by Wright and Panchepakesan (1969). This procedure has subsequently been labeled an unconditional maximum likelihood procedure (UCON)

and is the most widely used calibration technique currently available. Although the original article presents the technique in considerable detail, a more recent article (Wright & Douglas, 1977a) gives a clearer exposition.

The unconditional maximum likelihood procedure can best be summarized using the exponential form of the simple logistic model:

$$P\{x_{ij}\} = \frac{e^{x_{ij}(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} \tag{7}$$

where $\theta_i = \ln A_i$ and $b_j = -\ln E_j$. Using this equation, the likelihood of the entire matrix of responses of N persons to L items is given by

$$\Lambda = \frac{e^{\sum\sum_{ij} x_{ij}(\theta_i - b_j)}}{\prod\prod_{ij}(1 + e^{(\theta_i - b_j)})}. \tag{8}$$

Taking the logarithm of this equation simplifies matters substanitally, yielding

$$\lambda = \ln\Lambda = \sum_i r_i\theta_i - \sum_j s_j b_j - \sum\sum_{ij}\ln(1 + e^{(\theta_i - b_j)}) \tag{9}$$

where r_i is the raw score for Person i and s_j is the number of times Item j is answered correctly.

The first and second derivatives of Equation 9 are then computed with respect to θ and b. These derivatives are used, along with a sorting of the data into raw score groups to take advantage of the model's sufficient statistic properties to arrive at equations for finding the maximum of λ. A Newton-Raphson technique is used for this purpose, with iterations continuing until successive estimates become stable. A detailed description of this procedure can be found in Wright & Douglas (1977a).

Although the UCON procedure is the most widely used estimation technique for the simple logistic model, Andersen (1970) has shown that the unconditional approach yields inconsistent estimates. That is, as the sample size increases, the estimates do not approach the parameter values. Wright & Douglas (1977a, 1977b) have recently discussed this problem, and have shown that any bias induced is small and that it can easily be removed with a simple correction factor. The procedure does have the advantages of estimating ability and item parameters simultaneously and of being usable with lengthy tests and large samples.

A procedure that does produce consistent estimates of the parameters, the fully conditional procedures, has been developed by Andersen (1973) (FCON). This procedure uses the probability of a person's response string conditional on his raw score. The conditional probability is free of the ability parameter since the raw score is a sufficient statistic for ability. The actual procedure begins with the same exponential form for the simple logistic model as does the UCON procedure (Equation 7).

Since the responses to the items are assumed to be independent of one another, the probability of the response string for a person is given by

$$P\{[x_{ij}]\} = \prod_{j}^{L} P\{x_{ij}\} = \frac{e^{r_i\theta_i - \sum_{j}^{L} x_{ij}b_j}}{\prod_{j}^{L}(1 + e^{(\theta_i - b_j)})} \qquad (10)$$

where $[x_{ij}]$ denotes the vector of responses for person i.

The probability of a raw score r is the sum of the probabilities of the vectors of responses that yield that raw score.

$$P\{r\} = \sum_{[x_{ij}]}^{r} P\{[x_{ij}]\} = \frac{e^{r\theta_i}\gamma_r}{\prod_{i}^{L}(1 + e^{(\theta_i - b_j)})}$$

where $\gamma_r = \sum_{[x_{ij}]}^{r} e^{-\sum_{j}^{L} x_{ij}b_j}$ is the elementary symmetric function.

The conditional probability of the response string given the raw score is then

$$P\{[x_{ij}]|r\} = \frac{P\{[x_{ij}]\}}{P\{r\}} = \frac{e^{-\sum_{j}^{L} x_{ij}b_j}}{\gamma_r}$$

which does not contain the ability parameter θ.

The likelihood of the entire items-by-persons matrix can now be found, each person's vector conditional on his raw score

$$\Lambda = \frac{e^{-\sum_j s_j b_j}}{\prod_r \gamma_r^{n_r}} \qquad (11)$$

where s_j is the number of correct responses to Item j and n_r is the number of times raw score r was obtained. The logarithm of this value is used to simplify further computation

$$\lambda = \log \Lambda = -\sum_j s_j b_j - \sum_r n_r \log \gamma_r. \qquad (12)$$

Once this likelihood equation is obtained, solution for the item parameter follows much the same as for the UCON procedure. The first and second derivatives of Equation 12 are determined with respect to b_j and the Newton-Raphson iterative procedure is used to find the maximum value of λ.

Although FCON yields better estimates from a statistical point of view, the procedure suffers from computational difficulties due to the necessity of computing the elementary symmetric function. If a fifty item test were being calibrated using this procedure, computation of the elementary symmetric function for a raw score of twenty would require the sum of approximately 10^{13} terms. Even the fastest computer will be taxed by these computations. Therefore, the FCON procedure has been limited to application where the test being calibrated has less than fifteen items.

Wright and Douglas (1977a) have proposed a modification of one FCON procedure that attempts to solve the problem caused by the computation of the elementary symmetric functions. This procedure, called the incomplete conditional procedure (ICON), ignores selected symmetric functions, thereby improving computational efficiency. The resulting parameter estimates obtained using the ICON procedure are virtually the same as the FCON procedure, indicating that the revisions did not affect the accuracy of the method. However, despite the elimination of some of the symmetric functions, the procedure still becomes inaccurate if more than twenty to thirty items are used beacuse of accumulated roundoff errors. Thus the modifications to FCON only put off the point at which the procedure becomes unusable, and do not totally solve the problem.

Along with these five methods for estimating the parameter of the simple logistic model, two other more specialized methods should be mentioned. The first of these is a method labeled PROX by Wright and Douglas (1977b). This method was derived in an attempt to speed up the estimation process over the UCON method. The procedure starts with the initial estimates used by the UCON procedure, and then, using the assumption that the ability parameters are normally distributed, proceeds with a simplified iteration process to arrive at estimates. The simplified method shortens computation time by a factor of 40 for 50 to 60 item tests, but it is less accurate than UCON when extreme abilities are present or when the distribution of abilities is markedly skewed.

The second specialized method for estimation was developed by Meredith and Kearns (1973) and is called empirical Bayes point estimation. This procedure uses the expected value of the posterior distribution of the parameter of interest based on the raw score distribution to estimate the parameter. The procedure has been shown to be asymptotically optimal in the sense of having smaller average error variance and higher reliability than any other ability estimate when the sample approaches infinity. However, for sample less than 5000, the estimates tend to be unstable (Kearns and Meredith, 1975). This technique has not been extensively applied.

Three-Parameter Logistic Procedures

Due to the greater complexity of the three-parameter logistic model, the development of estimation procedures has taken longer than those for the simple logistic model. Fortunately, the logistic item characteristic curve closely approximates the normal ogive item characteristic curve (Birnbaum, 1968) allowing the adaptation of the previously developed normal ogive methodology to this mathematically more convenient model.

The first presentation of an estimation technique for the three parameter model was given in an appendix to an article by Lord (1968) concerning the analysis of the Verbal Scholastic Test. Except for the problems caused by the inclusion of a guessing parameter in the model, the method is similar to that used to obtain maximum likelihood estimates of the two-parameter normal ogive model (Lord, 1953).

The method begins with the three-parameter logistic equation for the probability of a correct response to an item

$$P_{ij} = P\{x_{ij} = 1\} = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \qquad (13)$$

where c_i is the guessing parameter for Item i, a_i is the discrimination parameter, and b_i is the difficulty parameter; θ_j is the ability parameter for person j; and D is a constant equal to 1.7 included to maximize the similarity of the model to the corresponding normal ogive model. Q_{ij} is defined as the probability of an incorrect response and is defined by $1 - P_{ij}$.

The first step in the estimation procedure is to determine the likelihood of the matrix of responses of the N persons to n items. This likelihood is given by

$$L = \prod_{i=1}^{n} \prod_{j=1}^{N} P_{ij}^{x_{ij}} Q_{ij}^{1-x_{ij}} \tag{14}$$

The logarithm of the likelihood is used for convenience.

$$\ln L = \sum_{i=1}^{n} \sum_{j=1}^{N} [x_{ij} \ln P_{ij} + (1 - x_{ij}) \ln Q_{ij}] \tag{15}$$

To determine the maximum of Equation 15, the derivative of the equation is determined relative to a_i, b_i, and θ_j. The guessing parameter, c_i, is not estimated using maximum likelihood at this point because the estimation procedure was found to be too unstable. Instead, the lower asymptote of the item characteristic value is used in the estimation equations for the other parameters. A sample of 100,000 cases was used in the original application study.

The three derivatives cannot be solved directly for zero because the individual parameters cannot be isolated as they could be in the simple logistic model. Instead, rough estimates of the item parameter are obtained using them. The resulting ability estimates are then used with the likelihood equations to obtain new item parameter estimates. The new item parameter estimates are then used to get new ability estimates, and so on. The two steps required to get estimates of the item and ability parameters are called a stage, and required about 15 minutes of computer time for the SAT analysis. Twenty stages were required for convergence to satisfactory values for that analysis.

Although the technique developed by Lord does yield usable results, several difficulties were encountered in its initial application. First, as mentioned above, the computer time required for the program is excessive. Approximately five hours of computer time were required for application to the Scholastic Aptitude Test data. The computer used (IBM 7044) was very slow compared to the current generation of computer, however, indicating that a substantial reduction would be obtained if it were run now. Second, tests used for calibration by this procedure should have at least 50 items and data should be available on at least

1,000 cases. If enough data is not available, the discrimination parameter may increase towards positive infinity. Third, occasionally ability parameters may go toward positive or negative infinity. Lord (1968) is not too disturbed by this fact because the result is expected whenever a perfect or zero score is obtained on the test. Fourth, despite large samples, the procedure may fail to converge in some circumstances. Lack of convergence may be caused by a single item and the procedure may converge when the item is removed. Sometimes, the item can be replaced after convergence has been achieved and a stable estimate can be obtained. Finally, the necessity of estimating the guessing parameter graphically from large samples makes the procedure impractical for many applications. Clearly, a more convenient procedure was needed to make the three-parameter logistic model more generally applicable.

In order to overcome many of these problems, an extended version of the procedure was made available in 1973 (Wingersky and Lord, 1973). This new version extended the maximum likelihood procedure to the guessing parameter, eliminating the need for extremely large samples to graphically estimate the lower asymptote of the item characteristic curve. However, Christoffersson (1975) has pointed out that estimating all of the item parameters and the ability parameters simultaneously is impossible unless constraints are placed on the parameters. He states that

> Intuitively, this seems impossible because the approach is equivalent to estimating factor loadings, factor scores and residual variance simultaneously in the case of interval measurement. This is not possible with the maximum likelihood method (Anderson and Rubin, 1956) unless some further conditions are imposed, such as assuming that the residual variances are pairwise the same.

To overcome this difficulty, numerous constraints have been built into the estimation program. Along with separating the estimation of ability and item parameters within a stage as was done in the 1968 version (Lord, 1968), the amount each parameter can change in each stage is restricted. This is done to reduce wild fluctuations in the estimates. Further, if a discrimination parameter exceeds a preset maximum value, the item is automatically removed from the analysis. Changes in the guessing parameters are severly restricted by the program since, in many cases, these parameters are poorly determined. Also, limits are placed on the minimum and maximum values allowed for the discrimination and guessing parameters.

With all of the constraints placed on the parameters, the procedure will converge on stable estimates if sufficient cases and test length are available. The number of items recommended has been reduced to 40 with this version, but the suggested minimum number of cases is still 1,000. Computation time has been reduced considerably. Time per stage on an

IBM 360/65 computer ranges from 70 to 180 seconds, with 30 to 40 stages required for convergence. Thus the computation time ranges from a half hour to two hours as compared to about five hours for the earlier version.

Along with improving the basic computational procedure of the maximum likelihood method, Lord also modified the procedure to recognize three modes of response: correct, incorrect, and omit (Lord, 1974). This was accomplished by allowing three item scores, 1 for a correct response, 0 for an incorrect response, and C for an omit response. The value of C used here is the reciprocal of the number of alternatives to the multiple choice item. The rationale for this scoring is that a person will only omit if he cannot guess at better than the chance level. Under those circumstances, the proportion of correct responses that would be expected if a person guessed would be equal to C, making this a reasonable level to use. The likelihood equation (Equation 14) is modified to reflect the scoring change yielding

$$L^* = \prod_{i=1}^{n} \prod_{j=1}^{N} P_{ij}^{v_{ij}} Q_{ij}^{i-v_{ij}} \qquad (15)$$

where $v_{ij} = 1$, 0, or C and the asterisk indicates the modified likelihood value. Lord (1974) points out that Equation 15 is not really a likelihood equation because of the change in scoring, but that it tends toward the same limit when the number of items is large. Also, Equation 15 yields smaller asymptotic sampling error than the maximum likelihood technique when the omitted responses are replaced by random responses. Thus, the modified equation was used in the 1973 version of the procedure (Wingersky and Lord, 1973).

The current step in the development of a procedure for estimation of the three-parameter logistic parameters is a revised edition of the 1973 program for increased efficiency. The revision, called LOGIST (Wood, Wingersky and Lord, 1976), has reduced the number of stages needed to reach convergence to 10 to 15. The greater efficiency was achieved by putting added constraints on the parameters. The ability parameters are restrained, where previously they were allowed to migrate anywhere between positive and negative infinity. Also, limits have been imposed on the discrimination and guessing parameters. The LOGIST program in its current and 1973 version is the most commonly used procedures for estimating the parameters of the three-parameter logistic model.

Because of the lengthy and expensive computation required by the maximum likelihood procedure, three other techniques have been developed to try to make parameter estimation more cost effective. The first, a graphic approximation method developed by Urry (1974), was originally designed for screening items. In using this method, the lower asymptote of the item characteristic curve must first be estimated from the item by total-score-minus-the-item regression. The c_i value found in this way

is used to select the proper nomograph for estimating the discrimination and difficulty parameters. The nomographs were generated from the theoretical relationship between the traditional population difficulty and discrimination values and the corresponding latent trait parameters. To use them, estimates of the population point biserial correlation and proportion correct are needed. These statistics are computed using traditional item analysis techniques. Once they are determined, they are entered into the ordinate and absissa of the nomographs, and the parameter estimates are read off a separate set of axes.

Urry (1974) has stated that estimates of the traditional item statistics based on a minimum of 2,000 cases are needed for good results, and the test should have at least 80 items and a KR-20 reliability of at least 0.90. When these conditions are met, the maximum likelihood and nomographic estimates are fairly comparable. In one study the a_i values were found to correlate 0.89 and b_i values 0.97. Urry was so impressed with these results that he feels the procedure may be used for final calibration of an item pool rather than as a mere screening device. He states that "It might well be that the heuristic estimates obtained through the present approximation method are to be preferred to maximum likelihood estimates where distortion of the estimates is artificially induced by the nature of the analysis [when low ability cases are dropped to improve LOGIST convergence]."

The second procedure developed to reduce the cost of estimating the latent trait parameters is called the ancillary estimation procedure (Urry, 1975). Although it is based on the normal ogive model, it is included here because the parameter estimates are very close to logistic values. This procedure is based on minimum chi-square estimation rather than the maximum likelihood used by most of the other procedures. The procedure involves two stages: the first stage uses raw scores as estimates of ability and the second stage uses Bayesian modal estimates of ability. The entire procedure can be summarized as follows.

First, initial parameter estimates are obtained for the item parameters by finding the minimum of a χ^2 variable given by

$$\chi^2_i = \sum_{j=0}^{m-1} \frac{(r_j - n_j P_i(j))^2}{n_j P_i(j) Q_i(j)} \qquad (16)$$

where χ^2_i is the result for Item i, r_j is the number of correct responses to Item i for those with raw score j, n_j is the number of cases obtaining a score of j, $P_i(j)$ is the probability of getting Item i correct for ability j based on the latent trait model, and $Q_i(j) = 1 - P_i(j)$. An iterative procedure is used to find the parameters a_i, b_i, and c_i, that are converted to ancillary estimates by correcting them using the Item i information functions (Lord & Novick, 1968). In effect, this correction

is an inverse weighting by the error of estimate. The purpose of the correction is to increase the efficiency of the estimates and to reduce the intercorrelations between the three parameters.

The ancillary item parameter estimates are then used to obtain Bayesian modal estimates of the ability parameters (Samejima, 1969). These estimates are obtained by finding the values of the ability parameters that maximize the following expression

$$B(\theta) = f(\theta)\prod_i P_i(\theta) \tag{17}$$

where $f(\theta)$ is the normal density function and $P_i(\theta)$ is the probability of a correct response to Item i as defined by the latent trait model. Once the new ability estimates are available, they are used to derive new minimum chi-square estimates of the item parameters. These new estimates are then again corrected using the information functions to get the final ancillary estimates.

The estimates obtained in this way were evaluated by Schmidt and Gugel (1975) to determine their effectiveness. They found that "Given at least 2,000 cases and 100 items of good but not unrealistically high quality, the procedure produces estimates that correlate highly with true parameter, show low root mean squares, and perform about as well when used in tailored testing as the true item parameter values." When the sample size drops as low as 500, the method may fail to converge.

By far the simplest of the procedures for calibrating items using the three parameter logistic model was presented by Jensema (1976). This procedure is designed mainly for screening items for further analysis and is not a substitute for maximum likelihood procedures. Jensema's procedure is based on the theoretical relationships that exist between the logistic parameters and traditional item analysis values as presented in Lord and Novick (1968).

In order to use this technique, the guessing parameter, c_i, must first be estimated from the lower asymptote of a plot of the item characteristic curve for the item. This value is used to adjust the proportion correct for the item using the formula

$$P_i = \frac{P_i' - c_i}{1 - c_i} \tag{18}$$

where P_i' is the estimated proportion correct for Item i and P_i is the corrected value. From the corrected value, the cutting point on the logistic distribution with that proportion above it can be obtained from

$$\gamma_i = \frac{1}{D}\ln\frac{1 - P_i}{P_i'} \tag{19}$$

where D = 1.7. This value, along with an estimate of the point-biserial correlation corrected for guessing, $\rho_{I\theta}$, is used to obtain the parameter estimate using the following formulas

$$a_i = \frac{\rho_{I\theta}}{\sqrt{1 - \rho_{I\theta}^2}} \tag{20}$$

and

$$b_i = \frac{\gamma_i}{\rho_{I\theta}}$$

where a_i is the discrimination parameter and b_i is the difficulty parameter.

Jensema has stated that reasonable estimates will be obtained using this method if three assumptions are met:

(a) reasonably good estimates of c_i are available,

(b) the proportion of the population passing Item i is an estimate of P_i, and

(c) the item-excluded total test score is a measure of true ability, θ.

The quality of the estimates obtained under these circumstances was checked using 48 simulation data-sets. The correlation between the true values and the estimates of discrimination and difficulty parameters for the procedure were found to be 0.798 and 0.963 respectively. The corresponding values for the maximum likelihood procedure were 0.863 and 0.971. Thus Jensema (1976) concludes that the estimates "which were inexpensive to calculate were surprisingly accurate." Sample sizes from 250 to 2,000 using 25 to 100 items were used for the study.

Summary All told, seven simple logistic and six three-parameter logistic procedures were identified for calibrating items. Other procedures available for two parameter models and normal ogive models are not included in the review. The seven simple logistic procedures include (a) Rasch's (1960) original graphic method for estimating parameters, (b) Brooks' (1964) modification of Rasch's method based on regression techniques, (c) the unconditional maximum likelihood procedure developed by Wright and Panchapakesan (1969), (d) the fully conditional maximum likelihood procedure developed by Andersen (1973), (e) the incomplete conditional developed by

Wright and Douglas (1977a), (f) the approximation procedure developed by Wright and Douglas (1977b), and (g) the empirical Bayes point estimation developed by Meredith and Kearns (1973).

Of these seven procedures, only three need be given serious consideration for use in calibrating items for tailored testing. The procedures developed by Rasch and Brooks have been largely supplanted by the newer procedures and are only of interest historically. The approximation procedure is designed for applications where limited computer resources are available, and therefore, should not be in consideration for the calibration studies considered here. The fully conditional procedure is too limiting in the small size of item pools and long computation time required. After eliminating these procedures from consideration, the unconditional procedure, incomplete conditional, and empirical Bayes procedure are left.

Of these three, the unconditional procedure seems to be the technique of choice for item calibration. Its only drawback is the slight inconsistency of its estimates, which can easily be corrected. The incomplete conditional procedure is limited by its constraints on test length, and no results are available on applications of the empirical Bayes procedure. Also, the empirical Bayes procedure is mainly concerned with ability estimation. Thus, for the purposes of the research reported here, the unconditional maximum likelihood procedure will be used.

The six three parameter logistic procedures include (a) thru (c) the three versions of the maximum likelihood procedure developed by Lord (1968), Wingersky & Lord (1973), and Wood, Wingersky & Lord (1976), (d) the nomographic procedure developed by Urry (1974), (e) the ancillary procedure developed by Urry (1975), and (f) the approximation procedure presented by Jensema (1976). Of these six procedures, the choice clearly falls between two. The early versions of the maximum likelihood procedure have been supplanted by an improved version making their use undesirable. Also, the nomographic and approximation techniques are clearly of lesser accuracy, leaving only the LOGIST procedure and Urry's ancillary estimation procedure. Of these two, the LOGIST procedure has been chosen for use here to avoid the assumption of a normal distribution of ability for the Bayesian modal estimates in the ancillary procedure. Although this assumption does not carry serious implications, the more generalizable procedure is preferred for the comparison to be conducted here.

Factors Affecting Item Calibration

Although the two calibration procedures selected for the research reported here were chosen because of their capabilities for arriving at accurate parameter estimates, neither of them will operate properly

under all circumstances. Both assume that the test being analyzed is unidimensional and that an adequate sample of observations is available. The simple logistic model also assumes that discrimination is constant for all items, and that guessing does not have an effect on the item responses. The research into the effects of these variables will now be summarized, with the goal of developing recommendations for the application of the methods.

Effects of Multivariate Test Data Despite the fact that the unidimensionality of the complete latent space is one of the basic assumptions of the latent trait models (Lord & Novick, 1968) and that the multidimensionality of tests is commonly used as an explanation for lack of fit (ie. Keifer & Bramble, 1974), only four studies could be identified that researched the effects of the factorial complexity of tests on item calibration. The lack of research is probably due to the common use of latent trait models with aptitude tests which can easily be constructed to contain a dominant first factor. When the first factor accounts for a moderate amount of the test variance, the latent trait models are felt to operate fairly well (Hambleton & Traub, 1973). However, with the application of the latent trait models to achievement tests, multidimensionality may become more of an issue.

The four studies that have been found in the literature search all deal with the robustness of the simple logistic model to violations of the unidimensional assumption. Only the study by Hambleton (1969) includes the two and three parameter models, and there they are secondary to the major thrust of the research. The first of the studies looking into the effects of violating the assumption was done by Hambleton (1969). In the study, he embedded either one or five items measuring a second factor in 15 or 30 item simulated unifactor tests. That is, if a 15 item test were used, one item would measure a second factor while the other 14 would measure the first factor. Goodness of fit of the simulated tests was then determined using a chi-square test and the number of rejected items was noted.

The results of this study showed that in all four cases the overall tests were rejected as fitting the simple logistic model on the basis of the chi-square tests. Also, with the increase in the number of items from the second factor, the chi-square values increased, and the number of items rejected by the model also increased. The conclusion drawn on the basis of these results was " . . . that the Rasch model is extremely sensitive to deviations from the assumption that the items of a test measure only one latent ability (Hambleton, 1969)."

In a second part to Hambleton's study, the one and two-parameter logistic models were applied to the Verbal and Mathematics Sections of the Ontario Scholastic Aptitude Test and the Verbal Section of the

Scholastic Aptitude Test. These tests were subjected to a principal factor analysis before the latent trait analyses and were found to have first factors accounting for 22.1%, 31.7% and 20.5% of the total variance respectively. Each of the tests was considered to have more than one factor, but the first factors were dominant.

Fit of the models to these tests was determined by generating theoretical frequency distributions from the item calibration results and comparing these distributions to the distributions obtained from the administration of the tests. A chi-square statistic was computed comparing each of the theoretical distributions with the actual distributions. In no case was the fit between the pairs of distributions good, a fact that was explained by the lack of unidimensional tests. The test with the largest first factor was found to have the smallest chi-square value, indicating the best fit. Also, the more general two-and three-parameter models yielded distributions that fit better than the one parameter model.

The second study that deals with robustness of the simple logistic model to violations of the univariate assumption was done by Reckase (1972). In this study one-, two- and three-factor simulated data-sets and one- and four-factor multiple-choice tests were analyzed using the model. As opposed to Hambleton's (1969) simulated tests which had one or five items from another factor embedded in them, the tests used by Reckase had equal numbers of items from each factor in the two-and three-factor simulated tests. The simulated tests were thirty items long and 1,000 cases were generated for each test.

The fit of the model to the tests and to each item was evaluated in the study using the chi-square test presented in Wright & Panchapakesan (1969). The results show that the one-factor data fit the model perfectly, the two-factor data did not fit at all, and the three-factor data fit moderately well. The same pattern occurred in terms of the number of items rejected; one item was rejected from the one-factor test, ten from the two-factor test, and five from the three-factor test. The loadings of the items on the factors were 0.90 in all cases and no guessing was present in the data indicating that all results were due to the factor structure of the test.

The results of the analysis on the three multiple-choice tests showed that none of the three tests fit the model well. However, the four-factor test had the poorest overall fit and had the greatest number of items rejected. The general lack of fit of these tests was probably due to the presence of guessing and unequal discrimination not included in the simulation data. The results of the two analyses show that multidimensionality does affect the fit of the simple-logistic model, but the fact that the three-factor simulation data fit reasonably well indicated that some robustness to multivariate effects may be present.

The other two studies that are related to the multidimensionality of test data do not control the factor structure of the tests as precisely as the first two. The study by Forbes & Ingebo (1975) initially calibrated a seventh grade mathematics test and then subjectively divided the test into three subtests labeled computations, problem application, and concepts. These subtests were then calibrated separately and the results compared to the overall calibration. In analyzing the data, it was argued that if the difficulty parameters differed only by a constant, there is no need to separate the items into homogeneous subtests for analysis. The results show that the items are ordered in the same way in the subtests and the total test, and that the calibrations yielded almost identical results. The authors concluded that the simple logistic model is sufficiently tolerant of violations of the "content homogeneity" assumption that the subtest breakdown is not necessary.

The final study to be described relating the factor structure of a test to the latent trait models was done by Ryan & Hamm (1976). This study attacked the problem from another direction by determining whether items selected to fit the simple logistic model would contain only one factor. Eight tests from a graduate research methods course were used for the study. Each of the tests was analyzed using the simple logistic model and the principal components factor analytic method. Items were selected from the tests which (a) were not rejected as fitting the simple logistic model, or (b) loaded highly on the first principal component. The selected items were again factor analyzed and size of the first factor was compared. The results showed that the size of the first factor was only slightly increased over the original test when items were selected using the simple logistic model, while factor analysis selected items had a substantially stronger first factor. Thus a simple logistic model cannot be looked on as a means of selecting homogeneous subtests. Checking the fit of at least the simple model to items is not a substitute for factor analysis.

The results of these studies yield few general conclusions. Clearly the Hambleton (1969) and Reckase (1972) studies show that the factor structure of tests affect the fit of the simple logistic model, but the effects on more complex models are lacking. The fact that the simple logistic model fit the three-factor test better than the two-factor test (Reckase, 1972) also suggests that the relation between factor structure and fit is not a direct one. Finally, the Ryan & Hamm (1976) study suggests that checking the fit of the models is not a substitute for factor analysis. Little about the effects of the violation of the unidimensional assumption is clarified by this literature review and some areas have not even been mentioned (i.e. the value of ability estimates obtained from multidimensional tests).

Effects of Sample Size Some information concerning the sample sizes required for stable calibration has already been presented under the review of calibration procedures. Wingersky and Lord (1973) suggest

a minimum of 1,000 cases for use with the LOGIST program. No controlled studies on the effects of sample size on calibration using the three parameter logistic model have been found.

Several studies have been done on the effects of sample size on the stability of the unconditional maximum likelihood estimates for the simple logistic model. Cypress (1972) calibrated a 90 item mathematics test using 1,200 normally distributed cases for use as a standard for comparison. She then calibrated the same test using independent samples of 1,200, 600, 300, 150, and 75 which also varied on seven levels of skewness. Thirty data-sets in all were used. The difficulty and ability parameter estimates were then compared to the estimates from the standard distribution.

In general, the study gave the expected results. As the sample size decreased, the standard error of the estimates increased. However, there is an interaction between the shape of the distribution and the similarity of the calibration to the standard distribution. The general conclusion of the study was that

> "If . . . intact groups reflect raw score distributions which are close to normal, results of the study indicate that groups as small as 75 may provide good ability estimates. Four groups consisting of 75 and 150 subjects ranked in the upper six when compared to the criterion group of 1,200 normally distributed tests scores. In fact, these four groups provided better estimates . . . than the group of 1,200 with low positive skew which ranked ninth."

A similar study by Forster (1976) compared calibration results from samples of 300, 200, 100, and 50 to calibration data from total samples of 1,478 and 1,808. Two tests were used for the study; an 81 item fourth grade mathematics test, and an 100 item eighth grade reading test. Correlations between the full sample and reduced sample parameter estimates were used as a basis for comparison. The largest drop in correlation was found between the samples of 100 and 50. On the basis of the results, the author concluded " . . . these results give us confidence in field testing with sample sizes of 150 to 200 to determine item difficulty calibrations with reasonable accuracy."

A third study (Tinsley & Dawis, 1975) compared the item and ability calibration results for ten pairs of intact groups ranging in size from 89 to 630. Four different tests with from 25 to 60 items were used for the study. The results show that if samples of over one-hundred are used, correlations in the high 80's or 90's can be expected between the item parameter estimates. With less than one-hundred, very low correlations were obtained. However, the correlations between ability estimates were found to be uniformly high, regardless of sample size.

The results of these three studies indicate that samples of a minimum of 100-150 are adequate for use of the simple logistic model. If a minimum of 1,000 is required for maximum likelihood estimates of the three-parameter model, the simple logistic model will have a clear advantage in cases where only small samples are available. A second implication of this research is that item parameters require more cases for accurate estimation than are required by ability parameters.

The Effects of Item Quality Item quality, for the purposes of this research, is defined in terms of the discrimination and guessing characteristics of the items. Poor quality items are those that are low in discriminating power or high in guessing. High quality items have the opposite characteristics. Classroom achievement testing often uses mediocre quality items for initial tests because of the short time allocated to test construction, and because the tests are often modified as the instructional process changes. Thus an important consideration in the evaluation of calibration procedures to be used for achievement tests is the degree to which the procedures are affected by this mediocre quality.

Three major studies have been found in the review of the literature that deal with the effects of discrimination and guessing on item calibration. This is not a complete list, since any study reporting the calibration of actual test data bears on the generalizability of the results, but the major findings present in the literature are presented in these studies. The first of the studies was done by Panchapakesan (1969). In the study, five, ten, and twenty item tests containing various numbers of "bad" items were analyzed using the simple logistic model. Bad items were items that were lower in discrimination than the majority of the items on the test. Simulated data with samples ranging from 100 to 2,000 were used.

The results of the study showed that items with discrimination values more than 0.2 below the average for the test can readily be detected as causing lack of fit. However, the model could be used adequately when the variation in the discrimination parameters of the items was not too extreme. Extreme in this case is defined as items with discrimination parameters deviating more than 0.2 from the average for the test.

Panchapakesan (1969) also looked into the effects of guessing on the simple logistic model, but not to the same extent as discrimination. Twenty item simulated tests using a sample of 5,000 cases were used for the study. Guessing levels of 0.5 and 0.2 were used to generate the simulation data. The results of the study indicated that guessing caused substantial errors in the calibration of hard items and in the ability estimates of low ability examinees. However, the effects on the easier items and on high ability estimates were negligible.

Panchapakesan recommended eliminating the hardest 25% of the items for calibration purposes when guessing is a factor and only accepting the ability estimates from the brighter individuals as being reasonably accurate.

The second study concerning item quality was done by Hambleton & Traub (1971). In the study, fifteen item tests with four ranges of discrimination parameters (0.0, 0.20, 0.40, 0.80) and three levels of guessing parameters (0.00, 0.10, 0.20) were compared on the basis of information level and relative efficienty using the one-, two-, and three-parameter logistic models. In general, the results came out as one would expect; the three-parameter model was found to be most informative, the two-parameter model next most informative, and the one-parameter model least informative. However, when guessing was present, the one-parameter model was better than the two-parameter model for low ability levels. When no guessing was assumed, the simple logistic model maintained high relative efficiency until the range of discrimination became large (0.80). On the basis of these results, the three-parameter model seems to be the recommended procedure, although sample size considerations did not enter into this study.

In the third study, Dinero & Haertel (1976) manipulated the variance of the discrimination parameters of 30 item simulated tests and compared the ability estimates from the simple logistic model with those from the two-parameter logistic model. Six different variances were used (0.0, 0.05, 0.10, 0.15, 0.20, 0.25) along with three different shapes for the discrimination parameter distributions (normal, uniform, and positively skewed). The uniform distribution was found to give the worst results overall. However, the lowest correlation between ability estimates was 0.8069 prompting the authors to conclude that the simple logistic model was robust to variations in discrimination.

In summary, the presence of guessing and excessive variation in the discrimination parameters affect the calibration of the one-parameter model to some extent, leading to recommendations to exclude low scoring cases and to select items on discrimination. The two parameter model seems to be affected more by guessing than the one-parameter model. Robustness to these factors is not a consideration with the three-parameter model since each of the parameters is estimated. Therefore, it has assumed the position of the standard by which the other techniques are judged.

Research Design

The research presented in this report is organized into three major components: (a) effects of multivariate test data, (b) effects of sample size, and (c) effects of item quality. Within each of these research components the two latent trait models, one-parameter and three-parameter logistic, are compared on their ability to estimate item and ability parameters. In addition, the two procedures are compared on the basis of cost and computer time required. However, before describing the specific analyses used to compare the procedures on these criteria, information concerning the data-sets and computer programs used in this research effort will be presented.

Data-Sets

The data-sets used for the research reported here are briefly described in Table 1 and the abbreviations used for them throughout this paper are presented. The data-sets are of three major types: (a) the results of the administration of standardized ability tests, (b) the results of the administration of college course final examinations, and (c) data generated to simulate tests with various factor structures. The standardized test results were acquired through the cooperation of the Missouri Statewide Testing Program. Results on the Missouri School and College Ability Tests were obtained for two school years, 1974-75 and 1975-76. The samples obtained were very large (57,800 in one case and 65,600 in the other), necessitating sampling from the total number to reduce the cost of the analyses. A sample size of 3,000 was selected since it was the maximum sample usable with the LOGIST program without modifications. This number was selected from the full sample using a systematic sampling procedure as there was no pattern to the original data. Both the Verbal and Quantitative subtests of each form were obtained from each sampling unit.

The final examinations from the undergraduate measurement course were obtained from five sections of A140: Introduction to Educational Measurement and Evaluation. This course covers basic measurement theory and practice for prospective teachers at the University of Missouri-Columbia. The data were collected from Fall 1975 to Spring 1977 from regular course examinations. Each of these examinations was constructed independently from a large item pool according to content specifications. All of the examinations, both the classroom and standardized tests, contained fifty multiple-choice items with four and five options respectively.

In order to gain greater control over the characteristics of the data, eight simulated test data-sets were produced. These were generated to match various factor loading matrices using the usual linear factor analysis model. The simulation procedure generated z-scores using a weighted sum of normal random numbers and then dichotomized them to yield the proportion

Table 1

Description of Data-Sets*

Test Name	Abbreviation	Sample Size	Description
1. Missouri School and College Ability Tests Verbal/1975	MSCATV5	3,087	Systematic sample from 57,800 cases from Missouri Statewide Testing Program 1974-1975. SCAT Series II Form 2B.
2. Missouri School and College Ability Tests Quantitative/1975	MSCATQ5	3,087	Systematic sample from 57,800 cases from Missouri Statewide Testing Program 1974-1975. SCAT Series II Form 2B.
3. Missouri School and College Ability Tests Verbal/1976	MSCATV6	3,126	Systematic sample from 65,600 cases from Missouri Statewide Testing Program 1975-1976. SCAT Series II Form 2B.
4. Missouri School and College Ability Tests Quantitative/1976	MSCATQ6	3,126	Systematic sample from 65,600 cases from Missouri Statewide Testing Program 1975-1976. SCAT Series II Form 2B.
5. Exam on Standardized Testing	ST1075	208	Undergraduate course final exam administered in October 1975.
6. Exam on Standardized Testing.	ST0576	181	Undergraduate course final exam administered in May 1976.
7. Exam on Standardized Testing	ST1076	176	Undergraduate course final exam administered in October 1976.

*All tests are 50 items in length.

Table 1 (Continued)

Description of Data-Sets

Test Name	Abbreviation	Sample Size	Description
8. Exam on Standardized Testing	ST3-577	312	Undergraduate course final exam administered to two sections of the course in March and May 1976.
9. One factor rectangular simulation data.	150AR	1,000	One factor with loadings of .9, rectangular distribution of difficulties.
10. Two factor normal simulation data.	250AN	1,000	Loadings of .9 and .0 randomly distributed on two factors, normal distribution of difficulties.
11. Two factor rectangular simulation data.	250AR	1,000	Loadings of .9 and .0 randomly distributed on two factors, rectangular distribution of difficulties.
12. Two factor .5 simulation data.	250A5	1,000	Loadings of .9 and .0 randomly distributed on two factors. All items .5 difficulty
13. Nine factor Spearman simulation data.	950ANS	1,000	One factor .7 loadings for all items. Eight factors .6 loadings randomly distributed over items. Normal disribution of difficulties
14. Nine factor independent .9 loading simulation data.	950AN9	1,000	Items randomly distributed to nine factors with .9 loadings. Normal distribution of difficulties.

Table 1 (Continued)

Description of Data-Sets

Test Name	Abbreviation	Sample Size	Description
15. Nine factor independent .3 loading simulation data.	950AN3	1,000	Items randomly distributed to nine factors with .3 loadings. Normal distribution of difficulties.
16. Five factor independent .7 loading simulation data.	550AN7	1,000	Items randomly distributed to five factors with .7 loadings. Normal distribution of difficulties.

of correct and incorrect responses specified by difficulty indices. These data-sets were produced without a guessing component, allowing a smaller sample size than the live testing data-sets. A sample of 1,000 cases, the minimum suggested by Wingersky and Lord (1973) for calibration, was generated for each of the eight simulated tests.

Four levels of factorial complexity were used in generating these data-sets: one-factor, two-factor, five-factor, and nine-factor. Of the eight data-sets, three were generated to have nine factors to match the empirically determined factor structure for the classroom tests. The size of the factor loadings and distribution of difficulties were also varied for the simulated tests. Normal, rectangular and constant distributions of difficulties were used, although no attempt was made to include all possible combinations. The distribution of difficulties referred to here is based on the proportion correct index.

Along with these data-sets, seven other samples were obtained for MSCATV6 to determine sample size effects. Systematic sampling was used, yielding samples of 2,929, 2,146, 1,494, 1,090, 748, 375, and 149. Care was taken to insure that no case occurred in more than one sample.

Computer Programs

Two computer programs are of major importance to this study. They are the maximum likelihood estimation procedures for the item and ability parameters for the one and three parameter logistic models. Since the comparisons between the two models are dependent upon the programs used for calibration,

it is important that the best available procedures be used. The programs were selected for this research on the basis of the review of the literature reported in the first part of this report. On that basis the unconditional-maximum likelihood procedure developed by Wright & Panchapakesan (1969) was selected for the one-parameter logistic model, and the quasi-maximum likelihood procedure for use with omitted responses developed by Wood, Wingersky, and Lord (1976) was selected for the three-parameter model. The actual program used for the one-parameter model was obtained from Jerry Durovic of the New York Civil Service Commission. The program was extensively modified by the author for greater efficiency and to correct some minor errors. The three parameter program was obtained from Marilyn Wingersky at the Educational Testing Service.

The program that generates the multivariate simulation data was written by the author for an earlier study using the random number generators from the International Mathematical & Statistical Libraries Package (1975). All other analyses were performed using the SPSS (Nie, Hull, Jenkins, Steinbrunner & Bent, 1975) and SAS (Barr, Goodnight, Sall, & Helwig, 1976) packages.

Effects of Multivariate Test Data

The purpose of the research reported here is to evaluate the one- and three-parameter logistic models for use in tailored achievement testing. Since the first step in setting up a tailored testing procedure is item calibration, this study first concentrates on that facet of the models - that is, determining item parameters. A complication in this matter is the fact that achievement tests tend to be multidimensional, violating the assumptions of both models. The evaluation of the item calibration procedures was therefore performed on the full set of 16 data-sets described earlier so that the effects of various factor structures could be ascertained.

The major quality desired in a calibration procedure is the ability to accurately estimate the item parameters so that the interaction of a person with the test is described with a minimum of error. The ability of each of the models to explain the interaction of the persons and the items was determined by comparing the predicted item response for a person with the actual item response. The predicted response is given by the probability of a correct response to the item for the ability level and is obtained from the appropriate model. The actual statistic used was obtained from the mean squared deviation of the obtained response from the expected response. The formula for the statistic is given by

$$MSD_i = \frac{\sum_{j=1}^{N}(u_{ij} - P_{ij})^2}{N} \qquad (21)$$

where u_{ij} is the response to Item i by Person j, P_{ij} is the probability of a correct response to Item i by Person j, and N is the number of people. The values of this statistic vary from zero for perfect prediction with a perfectly discriminating item; to .25 for an item with zero discrimination; to 1.0 for an item that predicts wrong responses when they are in fact correct. This statistic is used instead of the common comparison between theoretical and obtained item characteristic curves, because the latter fit statistic differs depending on the interval size used to approximate the empirical item characteristic curves.

Comparisons between the models were perform d on the MSD statistic using an ANOVA since the obtained values were approximately normally distributed. Thus, although the sampling distribution of this statistic was unknown, hypotheses could still be tested because only comparative information was of interest.

Along with the comparative information on item calibration from the two models, information on the factors controlling item calibration was desired. One statistic hypothesized to have some effect on the calibration procedures was the magnitude of the first eigenvalue of the tests. To determine if a relationship existed, the mean discrimination estimates from the 3PL model, the standard deviation of the difficulty estimates from the 3PL model, the standard deviation of easiness estimates from the 1PL model, the 1PL mean probability of fit, and the MSD statistic were plotted against the first eigenvalue. Correlations were also computed between the eigenvalues and these statistics across data-sets and the corresponding regression lines were obtained.

The interrelationships between the item parameters and the factor loadings used to generate the simulation data also yielded information about the test characteristics controlling the item calibration. To discover the relationships, the parameters were intercorrelated and later factor analyzed for summary purposes and to identify explanatory constructs. The live testing data and the simulated tests were analyzed separately using this procedure to determine if the simulated data findings were reproducible.

Once the quality of the item calibration data has been determined, the ability estimates based on the calibrated item pool become of interest. The questions of major importance concern the relationship of the ability estimates to the item responses, the relationship to outside criteria, and the relationship to factor scores. The relationships discovered among these variables will, in effect, define the construct being estimated by the latent trait models. Simple correlational techniques were used to determine the relationships between the various types of ability estimates. As with the study of item parameter estimates, all sixteen data-sets were used for these analyses. To

evaluate the relationship between the item responses and the ability estimates, the multiple correlations between the full set of item scores and the ability estimates for each of the two models were computed. The correlations were then compared using a t-test to determine which model explained more of the variance of the responses. Another analysis correlated the ability estimates from the two models with outside criterion measures available for the students from the undergraduate measurement courses. The available criterion included grades on other course exams. These correlations were also compared to determine which model gave ability estimates that were better predictors. Only three of the live testing data-sets could be included for this part of the study.

Effects of sample size

Another important question that has only been touched upon in the research literature is the sample size required for accurate estimation of parameters. To further explore the sample size limitations of the models, the seven subsamples of the MSCATV6 data-set were used. Parameter estimates were obtained from each of these samples and the results compared to the calibration based on 2,939 cases, using a squared deviation statistic. That is, for each of the item and ability parameters from the two models, the smaller sample estimates were subtracted from the large sample values, the difference squared, and the results summed. These estimates of squared deviations from the large sample estimates were then plotted against sample size in an attempt to identify the minimum sample size that yields adequate parameter estimates. Analysis of variance techniques were used to analyze the data.

Effect of item quality

Analyses were also performed on the data to determine what factors contributed to lack of fit of the models. To do this the MSD statistic presented earlier was correlated with the parameter estimates, traditional item analysis statistics, and factor analysis loadings. The purpose of the analysis was to discover what types of item should be eliminated from calibration studies. A similar analysis was done on the probability of fit obtained from the chi-square goodness of fit test used with the simple logistic model. These correlations were then factor anlayzed to summarize the results.

The final analysis performed on the two models was a comparison of computation cost and computer time. Although the results obtained from this analysis are computer specific, the proportions between the obtained values should generalize to other computer systems. These data will be required in future cost effectiveness studies of tailored testing.

Results

Summary statistics on the sixteen data-sets used in this study are presented in Table 2. Included for each test are: (a) the mean, (b) the standard deviation, (c) the KR-20 reliability, (d) the number of factors used to generate the data for the simulation data-sets, (e) the number of factors from the principal components analysis on phi-coefficients, (f) the first eigenvalue from the principal components analysis, (g) the number of factors from the principal factor analysis on phi-coefficients, (h) the first eigenvalue from the principal factor analysis, (i) the number of factors from the principal component analysis of tetrachoric correlations, (j) the first eigenvalue from the principal component analysis, (k) the sample size, (l) the CPU time for the simple logistic analysis, and (m) the CPU time for the three parameter logistic analysis. The three types of factor analysis were included in the study since the analysis of tetrachoric correlations sometimes yield non-Grammian matrices, and the analysis of phi-coefficients often yield difficulty factors. By using all three methods, it was hoped that the results of this study would be more generalizable.

Note that the principal factor analysis technique on phi-coefficients gave a fairly close approximation to the number of factors used to generate the data. Also, surprisingly, the KR-20 reliabilities are fairly high for all except the 950AN3 simulation data-set despite the fact that most of them are multidimensional. Other points of interest are that the classroom tests are easier than all of the others, and the three-parameter logistic program required substantially more computer time than the simple logistic program. The data reported in Table 1 will be used in many of the subsequent analyses.

Effects of Multivariate Test Data

To evaluate the effects of multivariate data on the two latent trait models, six analyses were performed: (a) the fit of the models to the data was determined, (b) the relationships between the first eigenvalue of the tests and various parameters of the tests were determined, (c) the relations between the item parameters and the item factor loadings were determined, (d) the relationships between the ability estimates and the factor scores were determined, (e) the relations between the item responses and ability estimates were determined, and (f) the relations between ability estimates and criterion measures were determined. The two procedures were compared in five of the six analyses. The sixth gives descriptive data only.

Goodness of fit of the models Deviations of the expected response derived from the models and the obtained response made by a person to an item were determined using Equation 21 given earlier. Using this equation, deviations

Table 2

Summary Statistics on the Sixteen Data-Sets

Statistic	Test Identifier						
	MSCATV5	MSCATQ5	MSCATV6	MSCATQ6	ST1075	ST0576	ST1076
Mean	29.02	28.52	29.14	28.67	35.00	35.00	34.00
Standard Deviation	9.57	9.92	9.22	9.40	4.10	5.00	5.30
KR-20	0.91	0.91	0.90	0.90	0.56	0.66	0.71
Expected # Factors	-	-	-	-	-	-	-
# Principal Component Factors[a]	8	8	9	9	21	21	20
First Eigenvalue	9.51	10.21	8.90	9.30	3.05	3.35	4.35
# Principal Factor Factors	2	3	2	3	9	9	9
First Eigenvalue	8.78	9.53	8.15	8.60	2.55	2.80	3.85
# Principal Component Tet Factors	8	8	8	9	22	21	20
First Eigenvalue	15.64	16.74	14.70	15.30	7.20	5.60	7.70
Sample Size	3087	3087	3126	3126	208	181	176
1PL CPU Time (Min)	0.36	0.36	0.37	0.37	0.51[b]	0.46[b]	0.45[b]
3PL CPU Time (Min)	4.19	4.80	4.49	5.22	1.20	1.11	1.10

[a]The number of factors for all factor analyses is based on the eigenvalue greater than 1.0 rule.

[b]These analyses were run off of cards and required scoring of tests. All other analyses were run off of tape and had been previously scored.

Table 2 (Continued)

Summary Statistics on the Sixteen Data-Sets

Statistic	Test Identifier								
	ST3-577	150AR	250AN	250AR	250A5	950ANS	950AN9	950AN3	550AN7
Mean	32.92	25.21	25.23	25.80	24.98	24.84	25.33	25.00	24.93
Standard Deviation	5.47	13.22	12.98	9.56	13.93	13.50	6.46	3.81	6.69
KR-20	0.69	0.97	0.95	0.93	0.96	0.96	0.74	0.22	0.76
Expected #. Factors	-	1	2	2	2	9	9	9	5
# Principal Component Factors[a]	21	4	4	7	2	9	9	22	6
First Eigenvalue	3.42	21.45	14.70	10.86	15.67	15.90	4.10	1.55	4.27
# Principal Factor Factors	5	3	3	4	2	9	9	22	6
First Eigenvalue	2.80	20.15	14.3	10.44	15.28	15.45	3.65	0.80	3.61
# Principal Component Tet Factors	21	4	2	6	2	9	9	22	6
First Eigenvalue	5.43	40.70	21.65	21.61	20.69	24.60	5.65	1.95	6.20
Sample Size	312	1000	1000	1000	1000	1000	1000	1000	1000
1PL CPU Time (Min)	0.15	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
3PL CPU Time (Min)	1.23	3.60	3.32	3.31	3.52	2.97	3.21	3.12	3.38

from fit for the models were determined for each item on each test. Thus 1,600 statistics were computed overall (16 tests X 2 models X 50 items). These were then used as the dependent measures in an analysis of variance to determine if the one-parameter or the three-parameter model fit the data better. The mean squared deviation from fit for the two models for the sixteen data-sets is given in Table 3 along with the analysis of variance summary table for the two way analysis design with repeated measures on one dimension.

The results show that the three-parameter model fits significantly better than the one-parameter model, although the difference in the overall means is only .004. However, for every data-set the average deviation from fit was smaller for the three-parameter model than for the one-parameter model. The deviations from fit were also found to be significantly different across tests. The 150AR data-set was fit best by the models as would be expected and the 950AN3 had the worst fit, also as expected. No interaction effect was found in the data.

To further rank the tests in terms of the fit of the models, the Newman-Keuls post hoc comparison procedure was used to determine if there were significant differences in the individual test means. The results of this analysis are presented at the bottom of Table 3. As can be seen from the results presented there, the 150AR data-set is fit by the models significantly better than any of the other tests. This is the one simulated test that meets all of the assumptions of both models. It contains only one factor, all of the items are equally discriminating, and no guessing is present.

The 250AR data-set has the next best fit for the models. It has two factors, a wide range of item difficulties, and no guessing. Although the fit for this test is significantly worse than 150AR, it is significantly better than all but one of the other tests. The majority of the other data-sets are fit about equally well by the two models. The best of these is ST1075, one of the classroom tests, and the worst is ST3-577, also one of the classroom tests. The standardized tests and the other two factor data-sets are included in this group.

At the poor fitting end of the continuum are three sets of simulation data: 550AN7, 950AN9, and 950AN3. All of these tests have a relatively large number of independent factors. Data-set 950AN3 is the worst fitting of the tests, having a MSD statistic very close to the value of .25 expected when all items have zero discrimination. This test has low loadings (.3) on the nine independent factors.

The trend of this analysis suggests that the multidimensionality of tests is a definite factor in the fit of the two models. The three-parameter logistic model handles this deviation from the assumptions significantly better than the one-parameter model, but the ordering of the effect is the same, as is shown by the lack of a significant interaction.

Table 3

Squared Deviations from the Two Models
for the Sixteen Data-Sets

Test	One Parameter Logistic	Three Parameter Logistic	Test Means
1. MSCATV5	.169	.166	.167
2. MSCATQ5	.164	.160	.162
3. MSCATV6	.169	.166	.167
4. MSCATQ6	.166	.161	.163
5. ST1075	.144	.138	.141
6. ST0576	.167	.165	.166
7. ST1076	.159	.154	.156
8. ST3-577	.184	.182	.183
9. 150AR	.068	.067	.068
10. 250AN	.162	.153	.158
11. 250AR	.122	.115	.118
12. 250A5	.185	.176	.180
13. 950ANS	.156	.156	.156
14. 950AN9	.211	.204	.208
15. 950AN3	.223	.222	.222
16. 550AN7	.210	.206	.208
Model Means	.166	.162	.164

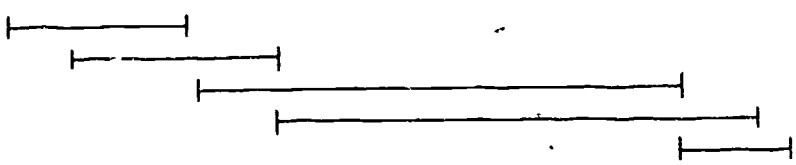
Anova Table

Source	Sum of Squares	d.f.	Mean Square	F	Significance
Tests	1.995	15	.133	31.667	.001
Items within tests	3.301	784	.004		
Models	.007	1	.007	14.684	.001
Tests X Models	.003	15	.0002	.414	
Models X Items within tests	.355	784	.0005		

Post Hoc Comparisons Using Newman-Keuls Test

Poor FIT	Test	Good FIT

15. 14. 16. 8. 12. 1. 3. 6. 4. 2. 10. 7. 13. 5. 11. 9.



Note: Those tests that are not underlined by the same line are significantly different from each other

Relationship to eigenvalues To further study the relationship between factorial complexity and goodness of fit, the first tetrachoric eigenvalue from the principal component analysis was plotted against the MSD statistic for each test. Figure 1 shows this relationship along with the regression line and correlation. The correlation between these two variables is -0.791 which is significant at the $p < .0005$ level, indicating that about 63% of the variation in fit can be accounted for by variation in the size of the first factor of a test.

An analysis of the scatter plot shown in Figure 1 shows that the three points that are below the regression line at the left of the graph are from three of the classroom tests. These tests were easier than the rest, suggesting that the difficulty of the tests might be a second variable explaining variation in the fit of the models to the test. To check this hypothesis, the multiple correlation among the average difficulty of the tests, the first eigenvalues, and the MSD statistic was computed, yielding a value of .935, a significant increase over the .791 obtained above. Figure 2 gives the scatter plot of the predicted MSD statistic, obtained from the average difficulty and the eigenvalue, and the actual MSD statistic. The three easy classroom tests have now moved closer to the expected regression line.

Figure 1

Relationship of the Average MSD Values to the First Eigenvalues

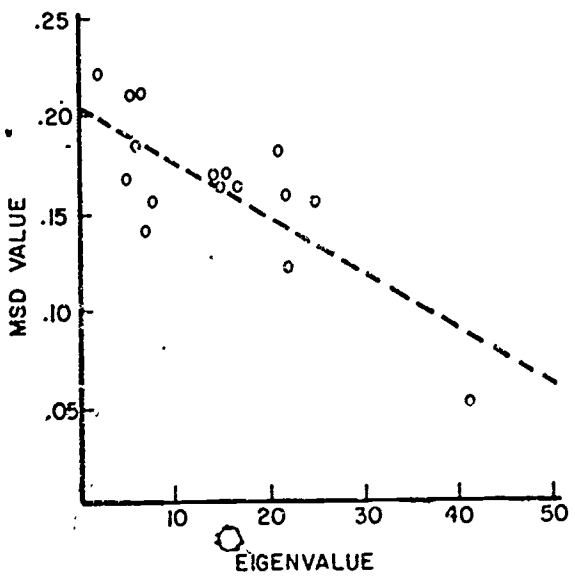
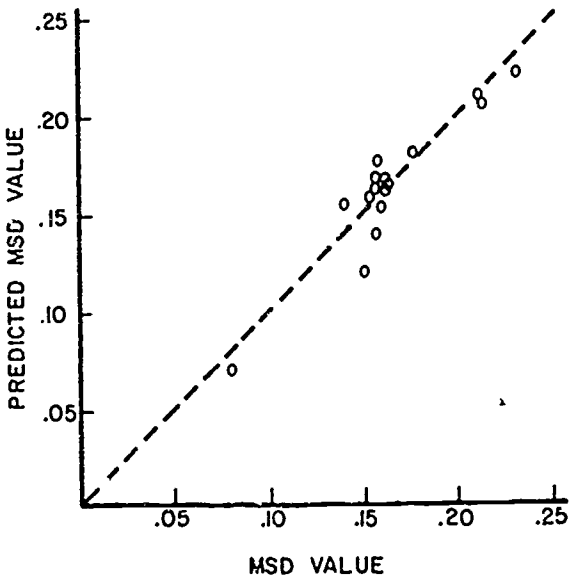


Figure 2

Relationship of the MSD Value Predicted from the First Eigenvalue and the Average Proportion Correct to the Obtained MSD Value



A second indication of the effects of multivariate data on the fit of the one-parameter logistic model is given by the relationship between the probability of fit obtained from the chi-square test in the Wright & Panchapakesan (1969) calibration program and the first eigenvalue. The plot of the probability of fit against the eigenvalue is given in Figure 3 along with the regression line. The relationship yields a correlation of .40, which is not significant at the .05 level. Further discussions of the usefulness of this probability of fit measure will be given later.

The effect of the size of the first eigenvalue on the operations of these two models was further analyzed to determine its relationship with several other statistics that define characteristics of the models. These statistics include the average 3PL discrimination parameter estimates, the standard deviation of the 3PL difficulty parameter estimates, and the standard deviation of the 1PL easiness parameter estimates.

The plot of the average discrimination parameter from the three-parameter logistic model against the first tetrachoric eigenvalue is given in Figure 4 along with the least squares regression line. Also included on the graph is the expected relationship between the eigenvalues and the average discrimination when all items have the same loading on the first factor. This relationship is given by the formula

$$\bar{a} = \frac{\sqrt{\frac{E}{N}}}{\sqrt{1 - \frac{E}{N}}} \qquad (22)$$

where E is the first eigenvalue, N is the number of items on the test, and $\bar{a}$ is the average discrimination parameter. This formula can be derived directly from that given in Lord & Novick (1968, Equation 16.10.7) by setting $P_g = \sqrt{\frac{E}{N}}$. This substitution assumes the normal ogive model rather than the logistic, as is used here. But since the two models yield very similar results, this equation should give approximation.

As can be seen from Figure 4, the first eigenvalues and the average discrimination have a strong relationship, yielding a correlation of .97. There is also a fairly close correspondence between the theoretical curve and the obtained data. None of these results are particularly exciting - they merely confirm theoretical expectations. However, they do give guidelines as to the required strength of the first factor required to obtain a particular average discrimination. For example, if an average discrimination of .8 is desired, Equation 22 yields a necessary first eigenvalue of 19.51 for a fifty item test, i.e. the first factor should account for 39% of the variance.

Figure 3

Relationship of the Chi Square Probability of Fit to the First Eigenvalue

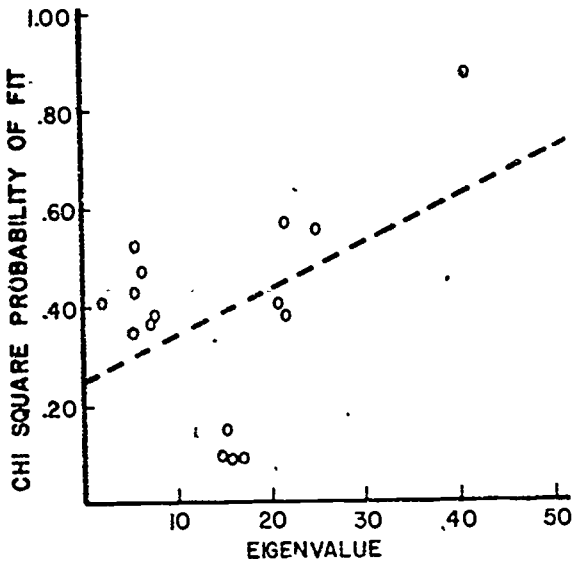


Figure 4

Relationship of the Average 3PL Discrimination to the First Eigenvalue

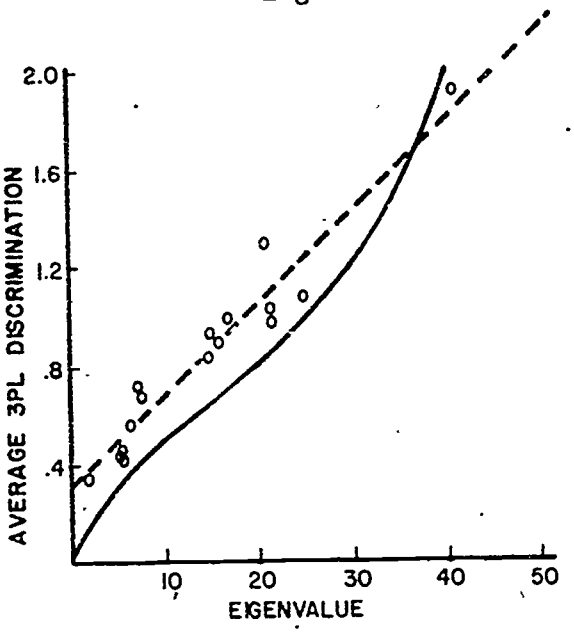


Figure 5

Relationship of the 3PL Difficulty Standard Deviation to the First Eigenvalues

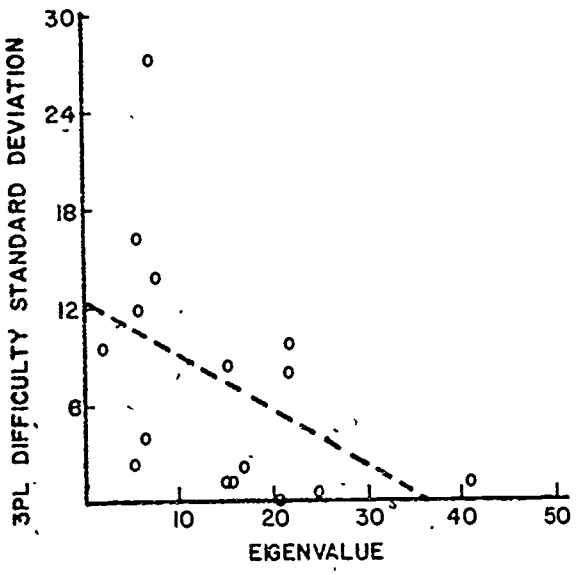
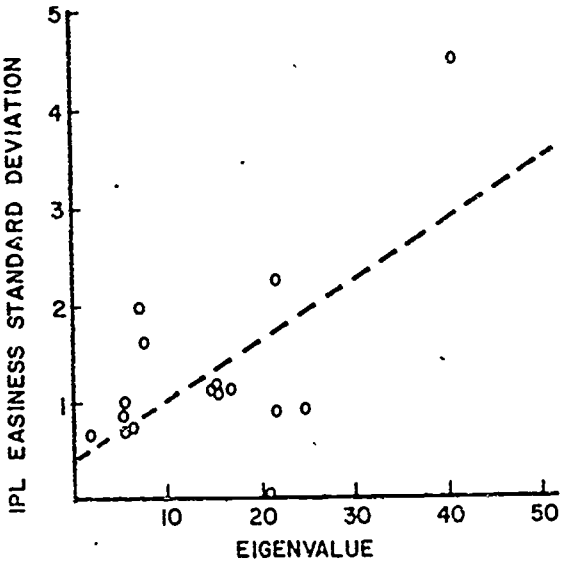


Figure 6

Relationship of the 1PL Easiness Standard Deviation to the First Eigenvalues



The plot of the first eigenvalues against the standard deviation of the difficulty parameters for the three-parameter logistic model is given in Figure 5. The standard deviation of the fit statistic is an indication of the stability of parameter estimates obtained by the calibration program. When convergence on estimates is poor, extreme values of the difficulty parameters generally appear in the calibration results, inflating the standard deviation. The correlation between the eigenvalues and the standard deviations is -.47 which is significant at the .05 level. The results generally show that when the eigenvalues are small, the results of the calibration tend to be unstable, giving larger values for the standard deviation of the difficulty parameters. The high variability in the lower eigenvalue range, however, indicates that caution is required in specifying any general rule. Several curvilinear functions were also checked for fit or this data, but none improved upon the simple linear regression line.

The scatter plot and regression line for the first eigenvalue against the standard deviation of the easiness parameters from the one-parameter logistic model are given in Figure 6. The standard deviation of the easiness values gives an indication of change of the ability scale of the model, usually brought about by differences in the average discrimination of the items (Baker, 1977). The correlation between the two variables is .62, indicating that as the size of the first eigenvalue increases, the spread of the parameter estimates increases. Thus, when the first eigenvalue is large, indicating high discrimination for the items, the items are widely spread or the ability scale shrinks. This is true even if the proportion correct for each item remains the same indicating that the size of the ability scale units has changed.

Relationship between item parameters The analysis up to this point has shown that there is a relationship between the factorial complexity of the data and the operation of the two latent trait models. These data do not show specifically what is being measured by the models under multivariate conditions. Therefore, several other analyses were done to determine the specific factor being evaluated by each of the models. These include a comparison between the factor loadings and the item statistics for the two models, and a comparison between the factor scores and the ability estimates.

The first data-set analyzed in this way is 250AN. This data-set was chosen for initial analysis since it is the simplest multifactor data-set available, allowing a clear indication of the relationship among the various parameters. Table 4 shows the loadings used to generate this data-set along with the 3PL discrimination parameter estimates, the 1PL probability of fit, the 1PL and 3PL MSD statistics, and the results of the factor analysis of this data. Table 5 gives the correlations between the variables presented in Table 4.

Table 4

Item Statistics for the 250AN Data-Set

Item	Theoretical Loadings I	Theoretical Loadings II	3PL a	3PL MSD	1PL Fit	1PL MSD	Varimax Factors I	Varimax Factors II	Principal Component Factor I
1	9	0	02	111	00	096	95	02	76
2	0	9	66	102	00	152	-07	90	50
3	0	9	133	90	01	155	00	88	54
4	9	0	08	191	17	145	89	08	75
5	9	0	11	204	02	155		03	74
6	0	9	186	84	48	148	10	90	64
7	0	9	10	89	42	160	03	91	59
8	9	0	10	218	22	165	91	02	73
9	9	0	192	224	00	173	91	-03	70
10	0	9	176	87	57	165	02	91	58
11	0	9	10	92	54	167	03	90	58
12	9	0	183	228	04	172	90	00	71
13	0	9	08	92	29	166	04	91	59
14	9	0	12	232	21	173	91	-01	71
15	9	0	165	234	52	168	92	02	73
16	0	9	178	99	92	176	-01	90	55
17	0	9	13	94	83	167	04	91	59
18	9	0	13	233	89	170	90	05	74
19	0	9	190	91	78	167	03	92	59
20	9	0	14	236	97	172	92	01	73
21	9	0	15	233	94	169	92	03	74
22	0	9	186	95	22	169	03	91	59
23	0	9	184	97	35	176	-01	91	55
24	9	0	14	235	67	173	90	02	72
25	0	9	197	93	25	173	-01	92	56
26	9	0	13	237	29	174	91	-01	71
27	9	0	12	238	20	175	92	-02	71
28	0	9	183	95	92	168	04	91	59
29	9	0	14	234	98	168	91	04	75
30	0	9	210	89	40	170	01	92	57
31	0	9	179	97	61	162	08	90	62
32	9	0	14	234	37	169	93	01	74
33	9	0	12	235	84	173	92	-01	72
34	0	9	210	91	93	169	02	91	58
35	9	0	15	230	22	164	91	05	75
36	0	9	186	95	70	173	-01	91	55
37	0	9	210	81	11	159	05	92	61
38	9	0	13	228	12	170	92	-00	73
39	0	9	210	84	15	163	00	93	58
40	9	0	12	226	12	174	91	-01	71
41	9	0	15	223	19	167	91	04	74
42	0	9	180	104	37	166	03	89	57
43	0	9	210	89	09	165	00	91	56
44	9	0	15	213	64	159	90	06	74
45	0	9	210	92	12	163	-02	90	54
46	9	0	15	199	21	157	92	01	73
47	9	0	14	191	20	151	90	03	73
48	0	9	193	78	26	140	04	89	58
49	0	9	210	81	01	143	-01	89	54
50	9	0	14	102	00	096	96	04	78

Note: All values presented without decimal points.

Table 5

Correlations between Factor Loadings, 1PL Fit, MSD Statistics, and 3PL Discrimination for 250AN

Variable	1	2	3	4	5	6	7	8	9
1. Factor 1[a]		-100	-97	93	-	-	100	-100	96
2. Factor 2[a]			97	-93	-	-	-100	100	-96
3. 3PL Disc				-90	-	-	-96	97	-90
4. 3PL MSD					-	29	91	-93	84
5. 1PL Fit						39	-	-	-
6. 1PL MSD							-	-	-
7. Varimax 1								-100	97
8. Varimax 2									-95
9. Principal Component 1									

Note: All correlations are presented without decimal points. Only significant correlations are presented.

[a]These factors are based on theoretical loadings.

In looking at Table 4, notice first the relationship between the 3PL discrimination parameter estimates and the theoretical loadings on Factor II. Without exception, the low discrimination estimates correspond to the zero loadings and the high discrimination values correspond to the high loadings. The correlation between these values is .97, confirming the subjective evaluation of the relationship. This relationship indicates that the 3PL model is differentiating among cases on the second factor of this simulated data-set. The empirically obtained loadings also yield this same result. Both verify the properties of the simulated data-set and reinforce the above findings.

The fit statistics also confirm the relationship between the 3PL model and the theoretical factor structure. The MSD statistic is consistently smaller for the items loading on the second factor than for those loading on the first factor. The correlation between the factor loadings and the MSD statistic for the 3PL model is -0.93, indicating the

strength of this relationship. Interestingly, neither the 1PL MSD nor the fit statistic are significantly correlated with the factor loadings. This indicates that the ability scale of the 1PL model does not seem to be related to any of the theoretical factors.

To further test this last hypothesis, the factor scores corresponding to the varimax factors and the first principal component were estimated and correlated with the ability estimates from the two latent trait models. These results are given in Table 6. The results presented here are somewhat of a surprise. Although the 3PL ability estimates are clearly more closely related to the second rotated factor than to the first, the correlation with the second factor is surprisingly low (.56). It is about the same size as the correlation with the first principal component and the raw scores. The 1PL ability estimates, on the other hand, correlate highly with the raw scores (the raw scores being a sufficient statistic for the ability estimates) and the first factor scores, and equally well with the two sets of rotated factor scores. The results are exactly what would be expected if the 1PL estimates were based on the sum of the scores on the two factors. On the basis of these results, it seems that the 3PL model is estimating the second factor, though rather poorly, while the 1PL model is estimating the sum of the two factors.

To confirm or deny that the 3PL model estimates one factor and the 1PL model estimates the sum of the factors, two other data-sets were analyzed: the 550AN7 data-set, and the 950AN9 data-set. The loadings, 3PL discrimination, and the fit statistics are given in Table 7 for the 550AN7 data-set and in Table 9 for the 950AN9 data-set. The correlations between the variables in these tables are presented in Tables 8 and 10 respectively. The correlations between the ability estimates and the factor scores are presented in Table 6 along with those from all of the other data-sets.

Notice first that, similar to the 250AN data-set, the items with high discrimination parameters correspond very closely to the items with .7 loadings on theoretical factor II. The first item is the only exception, probably due to the extreme difficulty of that item, making estimation of the parameters difficult. A similar relationship can be seen between the discrimination parameters and the second varimax factor loadings derived from tetrachoric correlations.

The correlations between the variables given in Table 8, reflect these subjective evaluations. The 3PL discrimination parameters correlate .91 with the second theoretical factor loadings and .92 with the second varimax factor. A smaller correlation is present with the first principal component, indicating that the first component is to some extent estimating the second varimax factor. The 3PL MSD statistic has a -.54 correlation with the second theoretical factor, supporting the overall conclusion.

Table 6

Correlation between Ability Estimates, Raw Scores, and Factors for the Sixteen Data-Sets

Data-set	Variable						
	Ability Estimate	Raw Score	3PL Ability	Phi Principal Component	Tet Principal Component	Tet Varimax 1	Tet Varimax 2
MSCATV5	3PL	97		98	98		
	1PL	99	96	97	97		
MSCATQ5	3PL	97		98	98		
	1PL	99	97	97	97		
MSCATV6	3PL	98		99	99		
	1PL	99	97	98	98		
MSCATQ6	3PL	97		98	98		
	1PL	99	96	97	97		
ST1075	3PL	83		89	32		
	1PL	99	85	89	29		
ST0576	3PL	88		91	87		
	1PL	99	90	93	88		
ST1076	3PL	89		94	91		
	1PL	98	90	88	86		
ST3577	3PL	95		98	98		
	1PL	99	95	97	97		
150AR	3PL	97		97	98		
	1PL	95	99	95	97		
250AN	3PL	59		59	56	29	56
	1PL	98	66	98	97	71	71
250AR	3PL	71		69	92		
	1PL	99	73	99	74		
250A5	3PL	82		56	62		
	1PL	98	83	76	83		
950ANS	3PL	93		93	94		
	1PL	98	96	98	98		
950AN9	3PL	62		82	67	74	
	1PL	99	62	72	72	44	
950AN3	3PL	71		36	41		
	1PL	100	71	25	33		
550AN7	3PL	70		46	36		64
	1PL	100	70	32	27		47

Note: All values presented without decimal points.

Table 7

Item Statistics for the 550AN7 Data-Set

Item	Theoretical Loadings II	3PL a	3PL MSD	1PL Fit	1PL MSD	Varimax Factor II	Principal Component Factor I
1	0	213	093	13	092	05	23
2	0	22	164	57	153	06	33
3	0	18	191	79	178	02	24
4	0	30	201	62	193	-02	-14
5	0	12	207	94	192	-01	32
6	0	35	212	26	199	-01	-08
7	0	18	223	34	212	05	31
8	0	09	227	49	209	-04	-48
9	7	162	161	08	214	77	48
10	7	213	149	14	213	75	46
11	7	196	159	95	222	75	43
12	7	195	149	47	215	75	49
13	0	11	242	90	220	01	30
14	0	05	241	08	227	-07	-52
15	0	09	241	24	224	08	-40
16	0	36	224	97	219	03	-06
17	0	04	239	50	224	-05	-52
18	0	32	229	24	225	00	-09
19	7	213	163	10	221	74	48
20	7	213	152	45	220	76	49
21	0	32	228	25	222	00	-11
22	0	13	246	12	229	04	35
23	7	164	168	07	230	74	44
24	0	08	248	88	230	01	-47
25	7	188	161	57	226	73	51
26	0	07	247	02	235	02	21
27	0	07	249	16	232	05	27
28	0	06	249	61	236	-02	-46
29	0	10	247	07	228	00	33
30	0	05	249	16	236	-02	19
31	0	34	221	51	221	01	-08
32	0	27	231	44	230	-05	-19
33	0	07	248	95	231	04	27
34	0	05	246	54	233	-05	23
35	7	166	175	36	229	74	49
36	7	169	172	70	224	74	47
37	0	05	243	10	231	-04	-49
38	0	03	240	68	223	-06	-49
39	0	03	241	10	226	-07	18
40	0	02	235	81	223	-03	17
41	0	34	194	42	201	01	-09
42	0	06	239	56	224	00	-46
43	0	09	228	24	215	-07	31
44	0	10	230	95	211	04	25
45	0	11	209	89	199	03	31
46	0	12	207	19	197	02	36
47	0	40	159	77	177	08	-02
48	0	36	150	69	167	04	-07
49	0	08	169	58	156	-01	31
50	0	10	096	92	093	-03	-48

Note: All values presented without decimal points.

Table 8

Correlations between Factor Loadings, 1PL Fit, MSD Statistics, and 3PL Discrimination for 550AN7

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Factor I[a]	-	-	-	-	-	-32	-	-	-	-	-29	-	99	-29	-83
2. Factor II[a]		-	-	-	-	91	-54	-	-	-	99	-	-28	-	56
3. Factor III[a]			-	-	-	-	-	-	-	-	-	-	-	99	-
4. Factor IV[a]				-	-	-	-	-	-	-	-	-	-	-	-
5. Factor V[a]					-	-	-	-	-	-	-	99	-	-	-
6. 3PL Disc							-71	-	-	-	92	-	-33	-	55
7. 3PL MSD								-	70	-	-57	-	-	-	-36
8. 1PL Fit									-	-	-	-	-	-	-
9. 1PL MSD										-	-	-	-	-	-
10. Varimax I											-	-	-	-	-30
11. Varimax II												-	-32	-	60
12. Varimax III													-	-	37
13. Varimax IV														-31	-86
14. Varimax V															-
15. Principal Component I															

Note: All values presented without decimals points. Only significant correlations are presented.

[a]These factors are based on theoretical loadings.

As with the 250AN data, the 1PL Fit for the 550AN7 data is not significantly related to any of the other statistics, suggesting that variations in discrimination or factor loadings are not a factor in lack of fit with this statistic. The only significant correlation with the 1PL MSD statistic is .70 with the 3PL MSD statistic, indicating that some of the error sources are the same, but that the common ones are not related to the factor structure, or to variation in discrimination.

The 950AN9 data further confirm these results. The high 3PL discrimination values correspond to the .9 loadings on Factor 9, except for the very difficult items. This observation also holds true for the first varimax factor. The correlational data in Table 10 gives similar results, yielding high positive correlations between 3PL discrimination and Theoretical Loadings IX and Varimax Factor I, and negative correlations between the 3PL MSD statistic and the same sets of loadings. The 1PL Fit has a barely significant correlation (.29) with Theoretical Factor I and a correlation of -.31 with the 1PL MSD statistic, again showing the lack of relationship between 1PL Fit and discrimination and factor structure.

Table 9

Item Statistics for the 950AN9 Data-Sets

Item	Theoretical Loading 9	3PL a	3PL MSD	1PL Fit	1PL MSD	Varimax Factor I	Principal Component Factor I
1	0	05	113	52	109	00	20
2	0	193	141	92	155	00	54
3	0	16	195	72	185	-01	51
4	0	01	1-8	21	191	00	-28
5	0	03	214	71	206	-04	05
6	0	193	197	70	201	01	58
7	0	193	192	70	200	-02	56
8	9	193	082	99	210	91	40
9	0	03	237	40	223	00	14
10	0	03	242	89	229	00	14
11	0	04	242	02	232	04	14
12	9	193	081	33	214	93	45
13	0	03	244	20	234	-05	01
14	0	06	242	64	229	-03	25
15	0	01	243	73	228	-01	-21
16	0	13	240	64	225	-01	15
17	0	24	232	43	225	05	53
18	0	05	247	08	234	-03	07
19	0	04	247	28	236	-02	03
20	9	193	086	29	217	92	47
21	0	06	248	25	236	02	05
22	0	24	233	24	223	06	52
23	0	06	248	99	232	-02	27
24	9	193	083	85	223	94	40
25	0	13	244	82	231	00	14
26	0	12	244	55	228	-02	19
27	9	193	087	43	222	93	42
28	0	15	242	08	227	01	57
29	0	04	247	40	234	-01	02
30	0	13	243	63	227	05	16
31	0	01	249	56	235	00	-21
32	0	02	243	39	229	02	29
33	0	13	244	96	226	-02	59
34	0	26	222	61	219	03	50
35	0	01	245	05	229	00	-21
36	0	02	239	19	228	00	12
37	0	07	243	24	231	04	04
38	0	15	237	31	224	06	59
39	0	09	233	36	215	04	32
40	0	15	230	62	217	01	17
41	0	23	217	18	212	02	51
42	0	06	232	58	220	04	09
43	0	01	229	37	213	02	-17
44	9	193	073	77	201	93	45
45	0	05	207	19	195	00	02
46	0	04	212	71	196	07	-14
47	0	22	187	60	186	01	45
48	0	06	179	88	171	07	03
49	0	10	166	74	155	02	31
50	0	07	095	81	091	00	28

Note: All values presented without decimal points.

Table 10

Correlations between Factor Loadings
1PL Fit, MSD Statistics, and 3PL Discrimination for 950AN9

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
1. Factor I[a]		-	-	-	-	-	-	-	-	-	-	29	-	-	-	-	99	-	-	-	-	-	-
2. Factor II[a]			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	90	-	-	-	-
3. Factor III[a]				-	-	-	-	-	-	32	-	-	-	-	99	-	-	-	-	-	-	-	52
4. Factor IV[a]					-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	99	-
5. Factor V[a]						-	-	-	-	-	-	-	-	-	-	100	-	-	-	-	-	-	41
6. Factor VI[a]							-	-	-	-	-	-	-	-	-	-	-	-	-	-	88	-	-
7. Factor VII[a]								-	-	-	-	-	-	-	-	-	-	91	-	-	-	-	-66
8. Factor VIII[a]									-	-	-	-	-	-	-	-	-	-	-	100	-	-	-
9. Factor IX[a]										78	-80	-	-	92	-	-	-	-	-	-	-	-	30
10. 3PL Disc.											-76	-	-	72	32	-	-	-	-	-	-	-	54
11. 3PL MSD												-	55	-75	-	-	-	-	-	-	-	-	-37
12. 1PL Fit													-31	-	-	-	-	-	-	-	-	-	-
13. 1PL MSD														-	-	-	-	-	-	-	-	-	-
14. Varimax I															-	-	-	-	-	-	-	-	29
15. Varimax II																-	-	-	-	-	-	-	55
16. Varimax III																	-	-	-	-	-	-	44
17. Varimax IV																		-	-	-	-	-	-
18. Varimax V																			-	-	-	-	66
19. Varimax VI																				-	-	-	-
20. Varimax VII																					-	-	-
21. Varimax VIII																						-	-28
22. Varimax IX																							-
23. Principal Factor I																							

Note: All values presented without decimal points. Only significant correlations are presented.

[a]These factors are based on theoretical loadings.

The correlations between the ability estimates and the factor scores presented in Table 6 show relationships similar to those for the 250AN data-set. The correlations between the 3PL ability estimates and the factor scores, corresponding to the various factor analytic solutions, confirm what was expected based on the previous analyses.

For the 550AN7 data-set, the correlation with the factor scores from the second varimax factor and the 3PL ability estimates is substantially higher than the correlation with the first principal component factor scores and the corresponding correlation with the 1PL ability estimates (.47). This latter correlation accounts for 22% of the variance while 20% would be expected if the 1PL ability estimates are based on the sum of the factors. A surprising finding for this data set is the .70 correlation between the raw scores and the 3PL estimates. No obvious explanation is available for this result.

The 950AN9 data-set gives similar results. The correlation of the 3PL estimates with the factor scores from the first varimax factor is much greater than that obtained using the 1PL estimates. The principal component factor scores have a slightly higher correlation with the 3PL ability estimates. This is probably due to the fact that these factor scores are based on several of the theoretical factors, as were the 3PL discrimination parameters (Factors 3 and 9). The varimax factor is a pure indication of theoretical factor 1.

In general, these simulation results indicate that when the data-sets are made up of equally weighted, independent factors, the 1PL model estimates the sum of the factors, while the 3PL model tends to estimate only one of the factors. This conclusion is a reasonable one based on the sufficient statistic properties of the 1PL estimates, and the factor analysis interpretations of the 3PL model (Christoffersen, 1975). However, most tests are not composed of equally weighted independent factors - instead they have a dominant factor with several smaller specific factors.

The 950ANS data-set simulates this type of test. Its first factor is large, and the other eight are relatively minor. In this case it does not make sense to correlate the item parameters with the theoretical loadings, since there is no variation in the loadings of the dominant factor. Therefore, the correlation with the factor scores was the only analysis performed in this case. Table 6 contains these correlations which are uniformly high for both models, indicating that both are estimating the first principal component.

The eight live testing data-sets also contain dominant first factors, although they are relatively small for the ST series, and therefore they also yield data bearing on this issue. In most of the cases (6 out of 8), the 3PL estimates correlate slightly higher than the 1PL estimates with the factor scores, while the 1PL estimates correlate higher with the

raw scores. In all cases, the correlations are substantial, the differences are small, and the estimates from the two models are highly related. These facts yield further evidence that the models are estimating the dominant factor when one is present.

Relationship to criterion measures Along with the above analyses that deal with what is being measured using the two latent trait models, two other analyses were performed that evaluate the two models relative to empirical criteria. The first analysis evaluates the relationship of the full set of test items to the ability estimates obtained from the models. This was done to determine the amount of variance in the items explained by the ability estimates. In order to determine this relationship, the multiple correlation between each of the ability estimates and the fifty items on each test was computed. These values are presented in Table 11 for the ability estimates from the two models correlated with the items from the sixteen data-sets. Note that all of the correlations with the 1PL ability estimates are extremely high, as they must be because of the sufficient statistic properties of that model. The multiple correlations are high for the 3PL ability estimates when a dominant factor is present, but drop when independent, equally weighted factors are present. This fact again supports the hypothesis that the 3PL model estimates a single factor since the correlation is reduced when items loading on other factors are present.

Table 11

Multiple Correlations Among
Ability Estimates and Test Items

	Ability Estimate		
Test	1PL	3PL	1PL-3PL
MSCATV5	.991	.983	.008
MSCATQ5	.998	.985	.003
MSCATV6	.993	.988	.005
MSCATQ6	.991	.983	.008
ST1075	.994	.944	.050
ST0576	.993	.952	.041
ST1076	.985	.967	.018
ST3-577	.996	.985	.011
150AR	.990	.997	-.007
250AN	.981	.677	.304
250AR	.991	.948	.043
250AS	.978	.839	.139
250ANS	.983	.949	.034
950AN9	.998	.852	.146
950AN3	.9998	.890	.1098
550AN7	.998	.866	.132
Mean	.9906	.9253	.07149

t = 3.705 p < .005

A related t-test was performed on the mean multiple correlations for the two ability estimates to determine if the observed differences are significant. The difference in the mean correlations of .07 is significant at beyond the .005 level indicating that the 3PL correlations are significantly lower.

The second analysis based on empirical data was a determination of the relationship between the ability estimates and outside criterion measures. This analysis showed which of the two models gave ability estimates with greater predictive power. The criteria used for this analysis were the first and second exam scores in an undergraduate measurement course. The correlations between the ability estimates and the two criterion measures are presented in Table 12. In all but one case, the 1PL ability estimates have higher correlations with the criteria than the 3PL estimates. However, in no case were the differences in correlations for the two models significant. One reason for the slightly lower correlations for the 3PL model could be the small sample size used in this analysis which would affect the 3PL model more than the 1PL model, causing unstable estimates.

Table 12

Correlations between Ability Estimates and Two Classroom Tests

Data Set	Ability Estimate	Test	
		Exam 1	Exam 2
ST1076	1PL	.555	.661
	3PL	.492	.599
ST0576	1PL	.409	.477
	3PL	.364	.483
ST1075	1PL	.558	.576
	3PL	.498	.535

Summary A total of six analyses was run on the sixteen data-sets to evaluate the effects of multivariate data on the two logistic latent trait models. These analyses dealt with the goodness of fit of the models to the data, the relationship of the parameter estimates to the size of the first factor of the test, the relationship of ability estimates to the factor scores, the relationship of ability estimates to item responses, and the relationship of ability estimates to outside criterion variables.

The MSD statistic, which was used as a measure of goodness of fit, showed that the 3PL model fit the data significantly better than the 1PL model. As the factorial complexity and unreliability of the tests increased, the fit decreased. This hypothesis was checked further by correlating the MSD statistic with the eigenvalues. The results indicated a strong negative relationship between the eigenvalues and the average MSD statistic. A follow up analysis indicated that the average difficulty of the test was a second major factor in the deviation of the model from fit. The 1PL probability of fit statistic was not significantly related to the size of the first eigenvalues.

Other variables that were found to be related to the size of the first eigenvalues were the average 3PL difficulty parameters (a measure of stability of estimation), and the standard deviation of the 1PL easiness parameters (an indication of change of scale). None of these relationships indicate new findings, but rather confirm theoretical expectations.

To determine what components in the tests were being estimated by the two models, the item parameter estimates were correlated with the theoretical and empirically obtained factor loadings. These analyses indicate that a single factor is estimated by the 3PL model while the 1PL model estimates the sum of the factor scores. The correlations of the factor scores with the ability estimates tend to confirm this finding and also show that when there is a dominant first factor, the two models estimate the same largest factor. If a number of equally powerful factors are present, there is no way to predict which factor will be estimated by the 3PL model.

The multiple correlations of the item response with the ability estimates show that the 1PL model has a significantly stronger relationship to the full set of items than the 3PL model. This finding is consistent with the contention that the 1PL model estimates the sum of the factors, therefore being affected by every item, and the 3PL model estimates a single factor, therefore only being affected by the items from that factor.

The two models did not differ significantly in their correlations with the outside criterion measures, although the 1PL model did have slightly larger values. Overall, the 3PL model fits the data better than the 1PL model, but this difference is not reflected in correlations with the outside variables. On the basis of these analyses, there is little to indicate the selection of one model over the other for the calibration of items for ability estimation when fifty item group tests are being used.

Effects of Sample Size

The effects of sample size on item calibration were determined by selecting systematic samples of various sizes from the Missouri Statewide Testing Data from the 1975-76 school year and then obtaining 1PL and 3PL item parameter estimates for each sample. Estimates of item parameters from each of the two calibration procedures were compared by computing the squared difference of each parameter estimate with the estimate from the largest sample. A one-way repeated measures analysis of variance was then performed using the squared difference values for the fifty items as the dependent measure with sample size as the independent variable to determine if the parameter estimates changed with sample size. The mean squared differences for each of the parameter estimates for the three 3PL item parameters and the one 1PL item parameter are given in Table 13 along with the analysis of variance results for the parameters.

The means of three of the four sets of item parameters give a similar pattern of results. The 2,997 sample has the smallest mean squared deviation, while the deviations tend to get larger with decreasing sample size. This relationship is strong for the 1PL easiness parameter and the 3PL discrimination parameter while the 3PL difficulty and guessing parameters show considerable variation. The analysis of variance results show significant differences in all cases except the 3PL difficulty parameters. In that case, although there are large differences in the means, the instability of the difficulty parameters causes large variation in the estimates resulting in a failure to reject.

The analysis of the variances of the squared deviations of the difficulty parameters showed extremely large differences. The ratio of the variances of the 1,090 sample to the 2,997 sample was 2,527, easily rejecting the hypothesis of homogeniety of variance using the F-max statistic ($F_{max} > 3.02$ needed for rejection). To compensate for this heterogeniety, a second ANOVA was performed on the 3PL difficulty parameter squared deviations after a logarithmic transformation was used ($Y = \log(x+1)$ (see Winer (1971), page 400). The ANOVA table for this second analysis is also given in Table 13. The revised analysis gave a significant F value indicating the presence of differences in the transformed mean squared deviation values for 3PL difficulty.

The purpose of this set of analyses was to determine at what point a decrease in sample size would adversely affect the results of item calibration. This question was addressed directly in a post hoc analysis performed using the ANOVA results. Using the mean squared deviation values for each sample size, the Newman-Keuls post hoc procedure was used to determine the largest sample that was significantly different from the mean squared deviation for the 2,997 sample. The results of these analyses are presented in Table 13. Samples that are not significantly different are underlined. Those that are different do not share the same underline.

Table 13

Comparison of Parameter Squared Deviations for tne Two Models by Sample Size

Sample Size	Parameter: 1PL Easiness	3PL Difficulty	3PL Discrimination	3PL Guessing
150	.0483	.1811(.1326)[a]	.2187	.0014
382	.0196	.1413(.0847)	.0973	.0009
763	.0063	.0272(.0258)	.0615	.0020
1090	.0063[.0064][b]	.1930(.0821)[.0260][b]	.0585[.0395][b]	.0009[.0011][b]
1525	.0055	.0299(.0263)	.0589	.0012
2197	.0047	.0138(.0135)	.0335	.0011
2997	.0041	.0166(.0162)	.0241	.0008

[a]Transformed means using log(x+1).

[b]Results from second sample.

ANOVA 1PL Easiness

Source	d.f.	SS	MS	F	P
Samples	6	.0791	.0132	18.17	<.0001
Error	294	.2133	.0007		

ANOVA 3PL Difficulty

Source	d.f.	SS	MS	F	P
Samples	6	2.009	.335	1.50	N.S.
Error	294	65.643	.223		

ANOVA 3PL Discrimination

Source	d.f.	SS	MS	F	P
Samples	6	1.303	.217	7.30	<.0001
Error	294	8.743	0.030		

ANOVA 3PL Guessing

Source	d.f.	SS	MS	F	P
Samples	6	0.000055	.000009	3.44	<.003
Error	294	0.000787	.000003		

Table 13(cont)

ANOVA Transformed 3PL Difficulty

Source	d.f.	SS	MS	F	P
Samples	6	0.627	.104	3.61	<.002
Error	294	8.506	0.029		

ANOVA Second Sample 3PL Difficulty

Source	d.f.	SS	MS	F	P
Samples	6	1.420	0.237	3.22	<.005
Error	294	21.630	0.074		

Post Hoc Comparisons

1PL Easiness

2997 2197 1525 1190 763 382 150

3PL Difficulty

2997 2197 763 1525 1190 382 150

3PL Discrimination

2997 2197 1190 1525 763 382 150

3PL Guessing

2-97 1190 382 2197 1525 150 763

The results of this analysis for the 1PL Easiness parameter are most easily interpreted. The largest sample significantly different from the 2,997 sample is the 382 sample while the 150 sample results are significantly greater than all of the others. These results suggest that there is little loss in calibration precision for the 1PL model until less than 763 cases are used. Although some loss is present for the 382 sample, the 150 sample is clearly inferior to all of the rest.

The analysis of the 3PL difficulty parameter estimates was more difficult to interpret. The order of the mean squared-deviations did not follow the sample sizes. The 1,090 sample had the largest squared deviations from the largest sample, followed by the 150 sample, the 382 sample, and the 1,525 sample. The 2,197 sample had the smallest squared deviation, followed closely by the 2,997 sample. The 763 sample had a much smaller squared deviation than was expected. Much of the surprising variation in this data was due to a few extreme estimates of the difficulty parameters. These occurred in cases where the items were very difficult or the discrimination was extremely high or low. The extreme values inflate the variance of the squared deviations causing the heterogeniety of variance mentioned above. The logarithmic transformation reduced the heterogeniety somewhat and also re-ordered the means slightly.

The results of the post hoc comparisons on the transformed 3PL difficulty data indicate that the 150 sample clearly deviates more than any other from the largest sample. None of the other mean squared deviations differ significantly from each other, even though some of the means are much larger than others. This can be attributed to the large difference in variances even after the logarithmic transformation. It should be noted that the smallest deviation for the 3PL difficulty parameter was about the same size as the 382 sample for the 1PL easiness parameter. Also, the asymptote on the mean squared deviation does not yet seem to have been reached for this parameter. Possibly, even larger samples are needed for stable calibration.

The 3PL discrimination parameter yielded fairly clear results. As the sample size decreased, the squared deviations increased. The 150 sample was significantly different from all of the other samples, while all of the others were not significantly different from each other. Samples of 382 or over, therefore, seem to estimate the discrimination parameters well, while samples of smaller size seem to result in inaccurate estimates. On the basis of this data, good estimates of discrimination parameters can be attained from much smaller samples than are required for the difficulty parameters.

The analysis of the guessing parameter squared deviations was less meaningful than the others because of the constraints placed on the parameter by the LOGIST program (Wood, Wingersky, & Lord, 1976). When very small samples were used, the guessing parameter estimates were not allowed to change at all, giving very small squared deviations. As the

sample sizes increased there were fewer constraints on the guessing parameter as the other parameter estimates became more stable. This resulted in greater squared deviations for the moderate samples. These deviations further decreased as the sample sizes increased. The ordering of the mean-squared deviations for the guessing parameter reflected this pattern. The 763 sample gave significantly larger deviations than the largest sample while the rest were not significantly different. Another interesting observation was that the squared deviations for this parameter were smaller than for all of the other parameters, showing the effect of the constraints.

Because the large squared deviations for the 1,090 sample were caused by a few extreme difficulty parameter estimates, a follow-up analysis was performed on a second sample of 1,088 to verify the results. The extreme values were not present in the analysis of this data, supporting a point of view that the extreme values were chance outliers. The squared deviations from this second sample are given in Table 13 along with the additional ANOVA results.

The mean squared deviation for this second sample for the 3PL difficulty was substantially lower than for the first sample (.0260 versus .1930) indicating the extreme variability of the sampling distribution of the squared deviations for the 3PL difficulty values and the possibility that the earlier sample contained several outliers. A second Newman-Keuls post hoc analysis gave the same results as the initial analysis. Also, heterogeniety of variance was still present in the analysis of the seven samples with the new data included. The re-analysis of the data after using the log-transformation resulted in no change in the results.

Due to the great variation in the 3PL difficulty values, the results of this study were not easily interpreted, indicating the need for further research. However, some general conclusions can be drawn from the data. The 1PL easiness parameters seem to have stabilized when the sample size is greater than 382. A sample somewhere between 382 and 763 is probably the lower limit required when using this model. The 3PL data are harder to interpret. The 3PL discrimination parameters seem to be moderately stable above the 150 sample, but the mean square deviations for the 3PL difficulty values are far from stable, with values for the 2,997 sample of about the same size as squared deviations for the 1PL easiness parameter for the 382 sample. Although these values are not on precisely the same scale, the values should be somewhat comparable. This result suggests that the 3PL difficulty parameters are just starting to stabilize. The heterogeniety of variance in the analysis of the difficulty parameters reduces its usefulness. However, the 150 sample is clearly worse than the rest. Overall the results suggest that substantially larger samples are required for the 3PL model. The guessing parameter does not enter into this discussion because of the numerous restrictions placed upon it.

Effects of item quality

The quality of items used in a test is indicated by the values of two parameters: the discrimination and guessing levels of the items. Items with high discrimination and low guessing parameters are items of high quality. To determine the effects of item quality on the calibration, the mean discrimination and guessing estimates for the eight live testing data-sets were computed and compared to mean values of other statistics available on the tests. The simulation data were not included in this analysis because no guessing component was included in the generation of the data, making it incompatible with the other data-sets.

The mean values of the parameters for the eight data-sets and the correlations of the means with the mean values for seven other test statistics are given in Tables 14 and 15 respectively. The mean 3PL discrimination statistics were found to be significantly related to three statistics; the 1PL Fit, the traditional difficulty (p), and the traditional discrimination ($r_{pt.bis}$). The correlation of the mean 3PL discrimination and the mean 1PL Fit statistic was -.86, indicating that tests with low 3PL discrimination tended to fit the 1PL model better than tests with high discrimination. This result was confounded by sample size, although sample size did not enter into the computation of fit. The MSCAT tests have high discrimination, low fit and have large samples; the ST series tests have lower discriminations, smaller samples, and fit the 1PL model better. The results tend to imply that mediocre tests fit the 1PL model better than highly discriminating tests, but caution must be used in generalizing this interpretation. Neither MSD fit measure was significantly related to mean 3PL discrimination.

Table 14

Means and Standard Deviations of the Guessing and Discrimination Parameter Estimates for Eight Data-Sets

Test Name	$\bar{X}_c$	S_c	$\bar{X}_a$	S_a
MSCATV5	.186	.036	.905	.410
MSCATQ5	.155	.043	.978	.484
MSCATV6	.153	.028	.840	.360
MSCATQ6	.155	.038	.959	.460
ST1075	.214	.052	.719	.682
ST0576	.158	.013	.466	.417
ST1076	.218	.027	.683	.630
ST3577	.160	.014	.447	.376

Table 15

Correlations Between Mean Discrimination and Guessing Parameters and Seven Other Variables

	1PL FIT	1PL MSD	3PL MSD	p	S_E	S_b	$r_{pt.bis}$
$\bar{X}_a$	-.86**	-.28	-.30	-.81**	-.60	-.33	.84**
$\bar{X}_c$	.42	-.66*	-.67*	.54	-.12	.62*	-.51

*p < .05
**p < .01

The second significant correlation with the mean discrimination estimates was with the mean traditional difficulty of the tests (-.81). As the tests became more discriminating, they tended to become more difficult. The average difficulty of the more discriminating tests was about .58 while those poorer in discrimination had an average difficulty of about .69. Again, these results were confounded by sample size, making the interpretation of the results unclea .

The third significant correlation with the mean 3PL discrimination occurred with the traditional discrimination index. This result was expected and showed the relationship between traditional and latent trait discrimination estimates.

Three variables were also significantly correlated with the mean 3PL guessing parameter: the 1PL MSD statistic, the 3PL MSD statistic, and the standard deviation of the 3PL difficulty estimates. The first two correlations imply that as guessing increases, fit to both models improves. This is the opposite of what was expected and it may be explained as an artifact of the 3PL calibration program. The estimates of the 3PL guessing parameter are only allowed to change from a pre-set value when good estimates are available for the difficulty and discrimination parameters. Good estimates are only likely to be available when the model closely fits the items. Thus high guessing values are only possible when the models closely fit the data.

The third correlation, between the mean 3PL guessing values and the standard deviation of the 3PL difficulty estimates, yielded the expected results. As guessing increased, the standard deviation of the difficulty estimates increased. As discussed earlier in this paper, the standard deviation of the difficulty indices is a measure of the stability of the calibration. Thus, as the amount of guessing on the items increased, the stability of the calibration tended to decrease (i.e. the standard deviation of the 3PL difficulty values increased).

Due to the many confounding variables in the above analysis, the results obtained were not easily interpreted. Therefore, a second analysis was performed within the test types to remove the confounding. This analysis intercorrelated the item statistics separately for the ST series tests and the MSCAT tests. Two hundred sets of item statistics were available for these two analyses. The obtained correlation matrices were then factor analyzed using the principal components technique and rotated to the varimax criterion. The factor analysis and rotation were done to summarize the relationship present between the item statistics and to determine what statistics were related to item quality. The rotated factor loading matrices for the MSCAT and ST series tests are presented in Table 16.

Although the MSCAT tests are of higher quality and the statistics are based on a larger sample than the ST series tests, the factor analysis of the ST series is easier to interpret because of greater variation in the item statistics, resulting in higher correlations and a clearer factor structure. Therefore, the results of the ST analysis will be discussed first and the MSCAT analysis will be used to reinforce the findings.

The principal components analysis of the ST series tests yielded four factors with eigenvalues greater than 1.0. These factors were rotated, yielding the factor loadings presented in Table 16. The first rotated factor has been labeled a discrimination factor with every discrimination statistic having a significant loading. The 1st Principal Factor and Component Loadings had the highest relationship to this factor and the 3PL discrimination values had the smallest significant loading. The magnitude of this latter loading was probably caused by instabilities due to the small sample size.

The second rotated factor was labeled a difficulty factor with high loadings from traditional difficulty and 1PL easiness. The 3PL difficulty statistic had a lower, but significant, negative loading. The negative sign was a result of the opposite scaling of the difficulty parameters. Three other statistics also loaded significantly on the factor; 3PL discrimination, 3PL guessing, and 1PL MSD. The presence of the discrimination parameter reflected the relationship between difficulty and discrimination discussed by Lord (1975). The guessing loading showed that as the easiness of the item increased, the size of the guessing parameter also increased. The 1PL MSD values tended to decrease with easier items, showing better fit.

The third factor was labeled a MSD fit factor. Both the 1PL and 3PL MSD fit statistics loaded highly along with 3PL discrimination. The loadings showed that with low discrimination the MSD statistic was large, as it should be, based on the discussion following Equation 22. Factor four was labeled a guessing factor, having high loadings on 3PL guessing and 1PL fit. The two loadings showed that when guessing was high 1PL Fit was low. This was the only factor with a significant loading for 1PL Fit leading to the conjecture that guessing was a major component in lack of fit using this statistic.

Table 16

Rotated Factor Loading Matrices
for the Item Statistics from the ST and MSCAT Data-Sets

Data-Sets	Variable	Factor I	II	III	IV
MSCAT	1st Principal Factor Loadings	.96	-.07	.14	.07
	1st Principal Component Loadings	.96	-.05	.14	.07
	1st Tetrachoric Principal Component Loadings	.94	-.20	.17	.04
	Traditional Discrimination	.69	-.10	.18	.32
	3PL Discrimination	.04	-.11	-.91	-.28
	Traditional Difficulty	.58	-.13	.67	-.22
	1PL Easiness	.58	-.21	.74	-.04
	3PL Difficulty	-.52	.18	-.01	.17
	1PL Fit	.14	.13	.12	.75
	1PL MSD	-.27	.92	-.07	.13
	3PL MSD	-.27	.92	-.04	.15
	3PL Guessing	.17	.64	.09	-.51
ST Series	1st Principal Factor Loadings	.97	.04	-.11	-.01
	1st Principal Component Loadings	.97	.04	-.09	-.01
	1st Tetrachoric Principal Component Loadings	.72	.05	.10	.23
	Traditional Discrimination	.95	-.06	.09	-.07
	3PL Discrimination	.45	-.57	-.59	-.22
	Traditional Difficulty	.09	.91	-.30	-.05
	1PL Easiness	.09	.93	-.24	-.02
	3PL Difficulty	.29	-.35	-.30	-.05
	1PL Fit	.26	.22	-.16	.64
	1PL MSD	.07	-.35	.88	-.22
	3PL MSD	.00	-.28	.93	-.13
	3PL Guessing	.17	.34	.07	-.76

Note: Significant values are underlined.

The MSCAT factor analysis yielded somewhat similar results, but with some confusion between the difficulty and discrimination factors. The discrimination indices had the highest loadings on the factor, with the exception of the 3PL discrimination, but the three difficulty indices also had significant loadings. It seemed that on this test, the easier items were more discriminating.

The second factor had high loadings with the 1PL and 3PL MSD statistics and with the 3PL guessing parameter. These loadings indicated a tendency toward poor fit when the items had high guessing parameters. This result was not found with the ST series tests.

The third factor was another mixture of difficulty and discrimination statistics. Traditional difficulty and the 1PL easiness statistics had moderate loadings on this factor, and 3PL discrimination had a large negative loading. The 3PL difficulty parameter, however, did not load on this factor. These results indicated that easy items were low in 3PL discrimination; a result that was directly opposite to those from Factor I. This indicated that the traditional and 1PL discrimination indices were not operating on the same component as the 3PL discrimination index.

The fourth factor showed the same pattern as the loadings for the ST series tests. The 3PL guessing parameter and 1PL Fit statistics gave the only significant loadings to this factor. Again, items with low guessing values had a high probability of fit on the 1PL model.

The effects of item quality shown by this analysis are threefold. First, guessing is the major factor in the lack of fit statistic for the 1PL model while discrimination seems to be unrelated to it. Second, guessing also seems to be related to the MSD statistics, but the results are not consistent across the tests. Third, the 3PL discrimination parameters are related to lack of fit in the ST series tests, but not for the MSCAT tests.

Calibration Costs

Since the earlier analyses in this report showed that the ability estimates were highly correlated when a dominant first factor was present in a test, there is little of a technical nature to use in selecting between these procedures when using them for ability estimation. Therefore, practical considerations become of importance in selecting a calibration procedure.

The major practical considerations of concern here are the cost of the calibrations and the storage requirements of the programs. The cost and CPU time for the GO-step of the two programs are given in Table 17 for the different data-set sizes used. The overall CPU time required for each data-set is given in Table 2. No significance tests

were required to determine that the LOGIST program costs substantially more, both in time and money, than the 1PL calibration program. On the average, the 3PL program cost 7.34 times as much and used 15.49 times as much CPU time for computation. The actual figures will probably not transfer directly to other computer systems, but the proportions should remain about the same. These figures were obtained from an IBM 370/168 computer system.

In terms of storage required, the 1PL program required 128K of core storage and one scratch unit for temporary storage. The 3PL program required 200K of core storage and two scratch units for temporary storage. Thus, the 3PL procedure was not only more expensive to run, but it also required more computer facilities.

One final note concerning the two procedures deals with the increase in cost as sample size increases. The cost of the 3PL procedure increased much faster than the 1PL procedure because ability estimates were obtained for each person. The 1PL procedure only obtained ability estimates for each score group. The increase in cost is reflected in the data presented in Table 17.

Table 17

Cost and Computer Time Required
for the UCON and LOGIST Procedures
for Various Sample Sizes

Procedure	Sample Size	Cost	Minutes CPU	Seconds Go-Step CPU
1PL	300	4.14	.15	3.59
	1000	4.57	.21	7.52
	3000	6.23	.36	16.89
3PL	300	15.36	1.23	38.33
	1000	37.31	3.30	162.72
	3000	48.91	4.68	245.19

Note: Sample sizes rounded to the nearest hundred.

Summary and Conclusions

The purpose of the research reported in this document was to evaluate the one- and three-parameter logistic models for use in calibrating items for tailored testing applications. However, several estimation

procedures have been developed for each of these models and specific procedures had to be selected before any comparisons could be made. To facilitate the selection process, a detailed review of the literature concerned with latent trait model calibration procedures was performed. Seven one-parameter and six three-parameter calibration procedures were identified and evaluated on the basis of statistical and practical characteristics. From the techniques reviewed, the procedure developed by Wright & Panchapakesan (1969) was selected for the one-parameter model, and the procedure developed by Wingersky, Wood, and Lord (1976) was selected for the three-parameter model. These procedures were felt to give the best combination of precision and practicality of those available.

The models, and their corresponding estimation procedures, were then compared on their ability to calibrate multivariate data, the sample size needed for calibration, the effects of item quality on the calibration, and the operational costs. Since the use of tailored testing with achievement tests is a long range goal of this research, the effects of multivariate data are of special importance. From the theoretical literature, it can be predicted that the three-parameter model would extract one factor from a set of items while the one-parameter model would be related to all of the factors present (Christofferson, 1975; Rasch, 1960). The analyses of the multivariate data-sets supported this point of view, showing that the three-parameter model computed item discrimination parameter estimates related to one factor, while the one-parameter estimates were related to the sum of the factors. However, when a relatively large first factor was present in the data, both procedures gave amazingly similar results. This finding was reflected mainly in the analyses of the ability estimates and the correlations of ability scores with outside criterion measures.

Despite the similarities found in the results of the procedures for ability estimates, the goodness of fit of the models to the data definitely showed the three-parameter model to be superior. A squared deviation statistic devised for this study was used for the goodness of fit analyses. This statistic gave a much better description of the operations of the procedures than the one-parameter probability of fit measure. This latter measure seemed to be unaffected by the multivariate nature of the data, or variations in discrimination, but was affected by guessing. These results indicate that the one-parameter probability of fit statistic is relatively uninformative concerning the fit of the model.

To further clarify the relation between the factorial complexity of tests and the calibration of item pools, a number of descriptive statistics were compared to the size of the first eigenvalue from the eighteen data-sets. The results showed that a strong relationship was present between the size of the eigenvalue and the average discrimination of the tests, the standard deviation of the difficulty and easiness parameters, and the squared deviation fit statistic. From the relationships, minimum recommendations can be made concerning the size of eigenvalue needed for a stable analysis.

If all other factors were equal, either of the two procedures would serve equally well for use with group administered tests. The three-parameter model accounted for more of the response variance with its item calibration procedure than did the one-parameter model, but the differences were small. No significant differences were found for ability estimates. However, all other factors were not found to be equal. The sample size required for stable calibration seemed to be much greater for the three parameter model, although the results were not totally conclusive. Also, the computation costs and the computer facilities required for the three-parameter procedure were substantially larger than those for the one-parameter model.

The research reported here deals with the use of calibration programs on data obtained from a group testing setting. Since item calibration is a necessary component of the tailored testing procedures based upon latent trait models, the evaluation is an important first step in the selection of a tailored testing procedure to be used for achievement measurement. However, the group nature of this data collection limits the generalizability of the results for individualized tailored testing. The one-parameter calibration procedure has been shown to give equivalent ability estimates at a lower cost than the three parameter procedure when basically unifactor group tests are used. However, the three-parameter procedure gives better fit to the item response data, a fact that may imply more usable item parameter estimates for tailored testing. Whether the better fit to the item responses will outweigh the higher cost of calibration can only be determined by a comparison of the usefulness of the ability estimates obtained from the procedures in live tailored testing. This comparison will be reported in the next technical report in this series.

REFERENCES

Andersen, E. B. Asymptotic properties of conditional maximum likelihood estimators. Journal of the Royal Statistical Society, 1970, 32, 283-301.

Andersen, E. B. Conditional inference for models for measuring. Copenhagen, Denmark: Mentalhygiejnisk Forlag, 1973.

Anderson, T. W. & Rubin, M. Statistical inference in factor analysis. In J. Neyanan (Ed), Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, (Vol. 5). Berkeley, California: University of California, 1956.

Baker, F. B. Advances in item analysis. Review of Educational Research, 1977, 47, 151-178

Barr, A. J., Goodnight, J. H., Sall, J. P., & Helwig, J. T. A user's guide to SAS76. Raleigh, North Carolina: SAS Institute, 1976.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Brooks, R. D. An empirical investigation of the Rasch ratio-scale model for item-difficulty indexes. (Doctoral dissertation, University of Iowa) Ann Arbor, Michigan: University Microfilms, 1964. (No. 65-434).

Christoffersson, A. Factor analysis of dichotomized variable. Psychometrika, 1975, 40, 5-32.

Cypress, B. K. The effects of diverse test score distribution characteristics on the estimation of the ability parameter of the Rasch measurement model. (Doctoral dissertation, The Florida State University) Ann Arbor, Michigan: University Microfilms, 1972. (No. 72-32, 756).

Dinero, T. & Haertel, E. A computer simulation investigating the applicability of the Rasch model with varying item discrimination. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, 1976.

Forbes, D. W. & Ingebo, G. S. An empirical test of the content homogeneity assumption involved in Rasch item calibration. Paper presented at the meeting of the American Educational Research Association, Washington, April, 1975.

Forster, F. Sample size and stable calibration. Paper presented at the meeting of the American Educational Research Association, San Francisco, 1976.

Hambleton, R. K. An empirical investigation of the Rasch test theory model. Unpublished doctoral dissertation, University of Toronto, 1969.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J.A. Developments in latent trait theory: a review of models, technical issues, and application. Paper presented at the joint meeting of the American Educational Research Association and the National Council on Measurement in Education, New York, 1977.

Hambleton, R. K. & Traub, R. E. Information curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 1971, 24, 273-281.

Hambleton, R. K. & Traub, R. E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

International Mathematical & Statistical Libraries, Inc. IMSL Library 1 Reference Manual, Houston: Author, 1975.

Jensema, C. J. The validity of Bayesian tailored testing. Educational and Psychological Measurement, 1974, 34, 757-766.

Jensema, C. A simple technique for estimating latent trait mental test parameters. Educational and Psychological Measurement, 1976, 36, 705-715.

Kearns, J. & Meredith, W. Methods for evaluating empirical Bayes point estimates of latent trait scores. Psychometrika, 1975, 40, 373-394.

Kifer, E. & Bramble, W. The calibration of a criterion-referenced test. Paper presented at the meeting of the American Educational Research Association, Chicago, April 1974.

Lord, F. M. An application of confidence intervals and maximum likelihood to the estimation of an examinee's ability. Psychometrika, 1953, 18, 57-76.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, 1970.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. The "ability" scale in item characteristic curve theory. Psychometrika, 1975, 40, 205-218.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley, 1968.

Marco, G. L. The application of item charactersitic curve methodology to practical testing problems. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, 1976.

Meredith, W. & Kearns, J. Empirical Bayes point estimates of latent trait scores without knowledge of the trait distribution. Psychometrika, 1973, 38, 533-554.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., Bent, D. H. Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1975.

Panchapakesan, N. The simple logistic model and mental measurement. Unpublished doctoral dissertation, University of Chicago, 1969.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Reckase, M. D. Development and application of a multivariate logistic latent trait model. (Doctoral dissertation, Syracuse University, 1972). Dissertation Abstracts International, 1973, 33. (University Microfilms No. 73-7762).

Reckase, M. D. An interactive computer program for tailored testing based on the one-parameter logistic model. Behavior Research Methods & Instrumentation, 1974, 6, 208-212.

Ryan, J. P. & Hamm, D. W. Practical procedures for increasing the reliability of classroom tests by using the Rasch model. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, 1976.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 1969, 34, Part 2.

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 1973, 38, 221-233.

Samejima, F. Behavior of the maximum likelihood estimate in a simulated tailored testing situation. Paper presented at the meeting of the Psychometric Society, Iowa City, 1975.

Schmidt, F. L. & Gugel, J. F. The Urry item parameter estimation technique: How effective? Paper presented at the meeting of the American Psychological Association, Chicago, 1975.

Tinsley, H. E. A. & Dawis, R. V. An investigation of the Rasch simple logistic model sample free item and test calibration. Educational and Psychological Measurement, 1975, 35, 325-339.

Urry, V. W. Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1974, 34, 253-269.

Urry, V. W. Ancillary estimators for the item parameters of mental test models. Paper presented at the meeting of the American Psychological Association, Chicago, August, 1975.

Wingersky, M. S. & Lord, F. M. A computer program for estimating examinee ability and item characteristic curve parameters when there are omitted responses. (ETS Research Memorandum RM-73-2). Princeton, New Jersey: Educational Testing Service, February 1973.

Wood, R. L., Wingersky, M. S. & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (ETS Research Memorandum RM-76-6). Princeton, New Jersey: Educational Testing Service, June 1976.

Woodcock, R. W. Woodcock Reading Mastery Tests. Circle Pines, Minnesota: American Guidance Service, 1973.

Wright, B. D. & Douglas, G. A. Best procedures for sample-free item analysis. Applied Psychological Measurement, 1977a, 1, 281-295.

Wright, B. D. & Douglas, G. A. Conditional versus unconditional procedures for sample-free item analysis. Educational and Psychological Measurement, 1977b, 37, 47-60.

Wright, B. D. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

DISTRIBUTION LIST

4 Dr. Marshall J. Farr, Director
Personnel & Training Research Programs
Office of Navy Research (Code 458)
Arlington VA 22217

1 ONR Branch Office
495 Summer Street
Boston MA 02210
Attn: Dr. James Lester

1 ONR Branch Office
1030 East Green Street
Pasadena CA 91101
Attn: Dr. Eugene Gloye

1 ONR Branch Office
536 S. Clark Street
Chicago IL 60605
Attn: Dr. Charles E. Davis

1 Dr. M. A. Bertin, Scientific Director
Office of Naval Research
Scientific Liaison Group/Tokyo
American Embassy
APO San Francisco 96503

1 Office of Naval Research
Code 200
Arlington VA 22217

6 Commanding Officer
Naval Research Laboratory
Code 2627
Washington DC 20390

1 LCDR Charles J. Theisen, Jr., MSC, USN
4024
Naval Air Development Center
Warminster PA 18974

1 Commanding Officer
U.S. Naval Amphibious School
Coronado CA 92155

1 CDR Paul D. Nelson, MSC, USN
Naval Medical R&D Command (Code 44)
National Naval Medical Center
Bethesda MD 20014

1 Commanding Officer
Naval Health Research Center
San Diego CA 92152
Attn: Library

1 Chairman, Leadership & Law Dept.
Div. of Professional Development
U.S. Naval Academy
Annapolis MD 21402

1 Scientific Advisor to the Chief
of Naval Personnel (Pers Or)
Naval Bureau of Personnel
Room 4410, Arlington Annex
Washington DC 20370

1 Dr. Jack R. Borsting
Provost & Academic Dean
U.S. Naval Postgraduate School
Monterey CA 93940

1 Mr. Maurice Callahan
NODAC (Code 2)
Dept of the Navy
Bldg. 2, Washington Navy Yard
(Anacostia)
Washington DC 20374

1 Office of Civilian Personnel
Code 342/02 WAP
Washington DC 20390
Attn: Dr. Richard J. Niehaus

1 Office of Civilian Personnel
Code 263
Washington DC 20390

1 Superintendent (Code 1424)
Naval Postgraduate School
Monterey CA 93940

1 Dr. H. M. West III
Deputy ADCNO for Civlian Planning
and Programming (Acting)
Room 2625, Arlington Annex
Washington DC 20370

1 Mr. George N. Graine
Naval Sea Systems Command
SEA 047C12
Washington DC 20362

1 Chief of Naval Technical Training
Naval Air Station Memphis (75)
Millington TX 38054
Attn: Dr. Norman J. Kerr

1 Principal Civilian Advisor
for Education and Training
Naval Training Command, Code 00A
Pensacola FL 32508
Attn: Dr. William L. Maloy

1 Dr. Alfred F. Smode, Director
Training Analysis & Evaluation Group
Department of the Navy
Orlando FL 32813

1 Chief of Naval Education and
Training Support (01A)
Pensacola FL 32509

1 Naval Undersea Center
Code 303
San Diego CA 92132
Attn: W. Gary Thomson

1 Navy Personnel R&D Center
Code 01
San Diego CA 92152

5 A. A. Sjoholm, Head, Technical Support
Navy Personnel R&D Center
Code 201
San Diego CA 92152

2 Navy Personnel R&D Center
Code 310
San Diego CA 92152
Attn: Dr. Martin F. Wiskoff

1 Dr. James McBride
Navy Personnel R&D Center
Code 310
San Diego CA 92152

1 Navy Personnel R&D Center
San Diego CA 92152
Attn: Library

1 Dr. Leonard Kroeker
Navy Personnel R&D Center
San Diego CA 92152

1 Dr. John Ford
Navy Personnel R&D Center
San Diego CA 92152

1 Dr. Richard A. Pollak
Academic Computing Center
U.S. Naval Academy
Annapolis MD 21402

1 Technical Director
U.S. Army Research Institute for the
Behavioral & Social Sciences
5001 Eisenhower Avenue
Alexandria VA 22333

1 Armed Forces Staff College
Norfolk VA 23511
Attn: Library

1 Commandant
U.S. Army Infantry School
Fort Benning GA 31905
Attn: ATSH-I-V-IT

1 Commandant
U.S. Army Institute of Administration
Attn: EA
Fort Benjamin Harrison IN 46216

1 Dr. Ralph Dusek
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria VA 22333

1 Dr. Joseph Ward
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria VA 22333

1 Dr. Ralph Canter
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria VA 22333

1 Dr. James L. Raney
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria VA 22333

1 Dr. Milton S. Katz, Chief
Individual Training & Performance
Evaluation Technical Area
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria VA 22333

1 HQ USAREUE & 7th Army
ODCSOPS
USAREUR Director of GED
APO New York 09403

1 DCDR, USAADMINCEN
Bldg. #1, A310
Attn: AT21-OED Library
Ft. Benjamin Harrison IN 46216

1 Dr. James Baker
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria VA 22333

1 Dr. Myron Fischl
Army Research Institute
5001 Eisenhower Avenue
Alexandria VA 22333

1 Research Branch
AFMPC/DPMYP
Randolph AFB TX 78148

1 AFHRL/AS (Dr. G. A. Eckstrand)
Wright-Patterson AFB
Ohio 45433

1 Dr. Marty Rockway (AFHRL/TT)
Lowry AFB
Colorado 80230

1 Instructional Technology Branch
AFHRL
Lowry AFB CO 80230

1 Dr. Alfred R. Fregly
AFOSR/NL, Building 410
Bolling AFB DC 20332

1 Air Force Human Resources Lab
AFHRL/PED
Brooks AFB TX 78235

1 Major Wayne S. Sellman
Chief, Personnel Testing
AFMPC/DPMYO
Randolph AFB TX 78148

1 Air University Library
AUL/LSE 76-443
Maxwell AFB AL 36112

1 Director, Office of Manpower
Utilization
HQ, Marine Corps (Code MPU)
BCB, Building 2009
Quantico VA 22134

1 Dr. A. L. Slafkosky
Scientific Advisor (Code RD-1)
HQ, U.S. Marine Corps
Washington DC 20380

1 Mr. Joseph J. Cowan, Chief
Psychological Research Branch (G-P-1/62)
U.S. Cost Guard Headquarters
Washington DC 20590

1 Dr. Harold F. O'Neil, Jr.
Advanced Research Projects Agency
Cybernetics Technology, Room 623
1400 Wilson Blvd.
Arlington VA 22209

1 Dr. Robert Young
Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington VA 22209

1 Mr. Frederick W. Suffa
Chief, Recruiting and Retention
Evaluation
Office of the Assistant Secretary
of Defense, M&RA
Room 3D970, Pentagon
Washington DC 20301

12 Defense Documentation Center
Cameron Station, Bldg. 5
Alexandria VA 22314
Attn: TC

1 Director, Management Information
Systems Office
OSD, M&RA
Room 3B917, The Pentagon
Washington DC 20301

1 Military Assistant for Human Resources
Office of the Director of Defense
Research & Engineering
Room 3D129, The Pentagon
Washington DC 20301

1 Dr. Lorraine D. Eyde
Personnel R&D Center
U.S. Civil Service Commission
1900 E. Street NW
Washington DC 20415

1 Dr. William Gorham, Director
Personnel R&D Center
U.S. Civil Service Commission
1900 E. Street NW
Washington DC 20415

1 Dr. Vern Urry
Personnel R&D Center
U.S. Civil Service Commission
1900 E. Street NW
Washington DC 20415

1 U.S. Civil Service Commission
Federal Office Building
Chicago Regional Staff Division
Regional Psychologist
230 S. Dearborn Street
Chicago IL 60604
Attn: C. S. Winiewicz

1 Dr. Joseph L. Young, Director
Memory & Cognitive Processes
National Science Foundation
Washington DC 20550

1 Dr. John R. Anderson
Dept. of Psychology
Yale University
New Haven CT 06520

1 Dr. Scarvia B. Anderson
Educational Testing Service
Suite 1040
3445 Peachtree Road NE
Atlanta GA 30326

1 Professor Earl A. Alluisi
Code 287
Dept. of Psychology
Old Dominion University
Norfolk VA 23508

1 Mr. Samuel Ball
Educational Testing Service
Princeton NJ 08540

1 Dr. Gerald V. Barrett
University of Akron
Dept. of Psychology
Akron OH 44325

1 Century Research Corporation
4113 Lee Highway
Arlington VA 22207

1 Dr. A. Charnes
BEB 203E
University of Texas
Austin TX 78712

1 Dr. Kenneth E. Clark
College of Arts & Sciences
University of Rochester
River Campus Station
Rochester NY 14627

1 Dr. Norman Cliff
Dept. of Psychology
University of Southern California
University Park
Los Angeles CA 90007

1 Dr. John J. Collins
Essex Corporation
201 N. Fairfax Street
Alexandria VA 22314

1 Dr. Rene V. Dawis
Dept. of Psychology
University of Minnesota
Minneapolis MN 55455

1 Dr. Ruth Day
Center for Advanced Study in
Behavioral Sciences
202 Junipero Serra Blvd.
Stanford CA 94305

1 Dr. John D. Carroll
Psychometric Lab
Davie Hall 013A
University of North Carolina
Chapel Hill NC 27514

1 Dr. Marvin D. Dunnette
Dept. of Psychology
University of Minnesota
Minneapolis MN 55455

1 ERIC Facility-Acquisitions
4833 Rugby Avenue
Bethesda MD 20014

1 Commanding Officer
Canadian Forces Personnel
Applied Research Unit
1107 Avenue Road
Toronto, Ontario, CANADA

1 Dr. Richard L. Ferguson
The American College Testing Program
P.O. Box 168
Iowa City IA 52240

1 Dr. Victor Fields
Dept. of Psychology
Montgomery College
Rockville MD 20850

1 Dr. Edwin A. Fleishman
Advanced Research Resources
Organization
8555 Sixteenth Street
Silver Spring MD 20910

1 Dr. John R. Frederiksen
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge MA 02138

1 Dr. Robert Glaser, Co-Director
University of Pittsburgh
3939 O'Hara Street
Pittsburgh PA 15213

1 Dr. Richard S. Hatch
Decision Systems Assoc., Inc.
5640 Nicholson Lane
Rockville MD 20852

1 Dr. M. D. Havron
Human Sciences Research, Inc.
7710 Old Spring House Road
West Gate Industrial Park
McLean VA 22101

1 Dr. Duncan Hansen
School of Education
Memphis State University
Memphis TN 38118

1 CDR Mercer
CNET Liaison Officer
AFHRL/Flying Training Div.
Williams AFB AZ 85224

1 HumRRO/Western Division
27857 Berwick Drive
Carmel CA 93921

1 Dr. Lawrence B. Johnson
Lawrence Johnson & Associates, Inc.
Suite 502
2001 S. Street NW
Washington DC 20009

1 Dr. Steven W. Keele
Dept. of Psychology
University of Oregon
Eugene OR 97403

1 Dr. Alma E. Lantz
University of Denver
Denver Research Institute
Industrial Economics Division
Denver CO 80210

1 Dr. Frederick M. Lord
Educational Testing Service
Princeton NJ 08540

1 Mr. Brian McNally
Educational Testing Service
Princeton NJ 08540

1 Dr. Robert R. Mackie
Human Factors Research, Inc.
6780 Corton Drive
Santa Barbara Research Park
Goleta CA 93017

1 Mr. Edmond Marks
304 Grange Bldg.
Pennsylvania State University
University Park PA 16802

1 Dr. Leo Munday
Houghton Mifflin Co.
P.O. Bos 1970
Iowa City IA 52240

1 Richard T. Mowday
College of Business Administration
University of Oregon
Eugene OR 97403

1 Dr. Donald A. Norman
Dept. of Psychology C-009
University of California, San Diego
LaJolla CA 92093

1 Mr. Luigi Petrullo
2431 N. Edgewood Street
Arlington VA 22207

1 Dr. Steven M. Pine
N 660 Elliott Hall
University of Minnesota
75 East River Road
Minneapolis MN 55455

1 Dr. Lyman W. Porter, Dean
Graduate School of Administration
University of California
Irvine CA 92717

1 Dr. Diane M. Ramsey-Klee
R-K Research & System Design
3947 Ridgemont Drive
Malibu CA 90265

1 R.Dir. M. Rauch
P II 4
Bundesministerium der Verteidigung
Postfach 161
53 Bonn 1, GERMANY

1 Dr. Joseph W. Rigney
University of So. California
Behavioral Technology Laboratories
3717 South Grand
Los Angeles CA 90007

1 Dr. Andrew M. Rose
American Institutes for Research
1055 Thomas Jefferson St. NW
Washingtc DC 20007

1 Dr. Leonard L. Rosenbaum, Chariman
Dept. of Psychology
Montgomery College
Rockville MD 20850

1 Lr. Richard Snow
Stanford University
School of Education
Stanford CA 94305

1 Mr. Dennis J. Sullivan
c/o Canyon Research Group, Inc.
32107 Lindero Canyon Road
Westlake Village CA 91360

1 Dr. Benton J. Underwood
Dept. of Psychology
Northwestern University
Evanston IL 60201

2 Dr. David J. Weiss
Dept. of Psychology
N660 Elliott Hall
University of Minnesota
Minneapolis MN 55455

1 Dr. Keith Wescourt
Dept. of Psychology
Stanford University
Stanford CA 94305

1 Dr. Anita West
Denver Research Institute
University of Denver
Denver CO 80201

1 Dr. Earl Hunt
Dept. of Psychology
University of Washington
Seattle WA 98105

1 Dr. Thomas G. Sticht
Assoc. Director, Basic Skills
National Institute of Education
1200 19th Street NW
Washington DC 20208

1 Prof. Fumiko Samejima
Dept. of Psychology
Austin Peay Hall 304C
University of Tennessee
Knoxville TN 37916

1 Dr. John Wannous
Dept. of Management
Michigan State University
East Lansing MI 48823

1 Dr. Frank Pratzner
The Center for Vocational Education
Ohio State University
1960 Kenny Road
Columbus OH 43210

1 Dr. Meredith Crawford
5605 Montgomery Street
Chevy Chase MD 20015

1 Dr. Nicholas A. Bond
Dept. of Psychology
Sacramento State College
6000 Jay Street
Sacramento CA 95819

1 Dr. James Greeno
Learning R&D Center
University of Pittsburgh
3939 O'Hara Street
Pittsburgh PA 15213

1 Dr. Frederick Hayes-Roth
The Rand Corporation
1700 Main Street
Santa Monica CA 90406

1 Dr. Robert Sternberg
Dept. of Psychology
Yale University
Box 11A, Yale Station
New Haven CT 06520

1 Dr. Walter Schneider
Dept. of Psychology
University of Illinois
Champaign IL 61820

1 Dr. Richard B. Millward
Dept. of Psychology
Hunter Lab
Brown University
Providence RI 82912